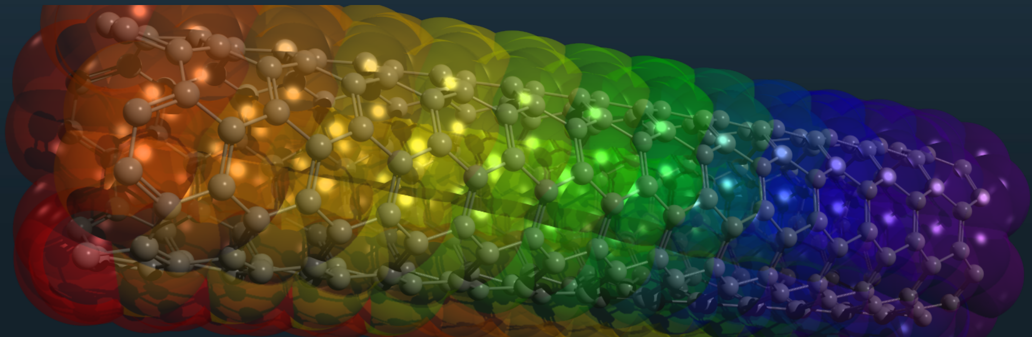
Chemical Databases and Open Chemistry on the Desktop

5th Meeting on US Government Chemical Databases & Open Chemistry
August 25, 2011

Dr. Marcus D. Hanwell
marcus.hanwell@kitware.com



Outline

- Background
- Opening up chemistry
- Workflows in computational chemistry
- Avogadro – chemical editor
- Databases on the desktop
- Quixote
- HPC resource integration
- Advanced visualization

My Background

- Ph.D. (Physics) – University of Sheffield
- Google Summer of Code – Avogadro
- Postdoc (Chemistry) – University of Pittsburgh
- R&D engineer – Kitware, Inc
- Passionate about physics, chemistry, and the growing need to improve computational tools
- See the need for powerful open source, cross platform frameworks and applications
- Develop(ed): Gentoo, KDE, Kalzium, Avogadro, Open Babel, VTK, ParaView, Titan, CMake

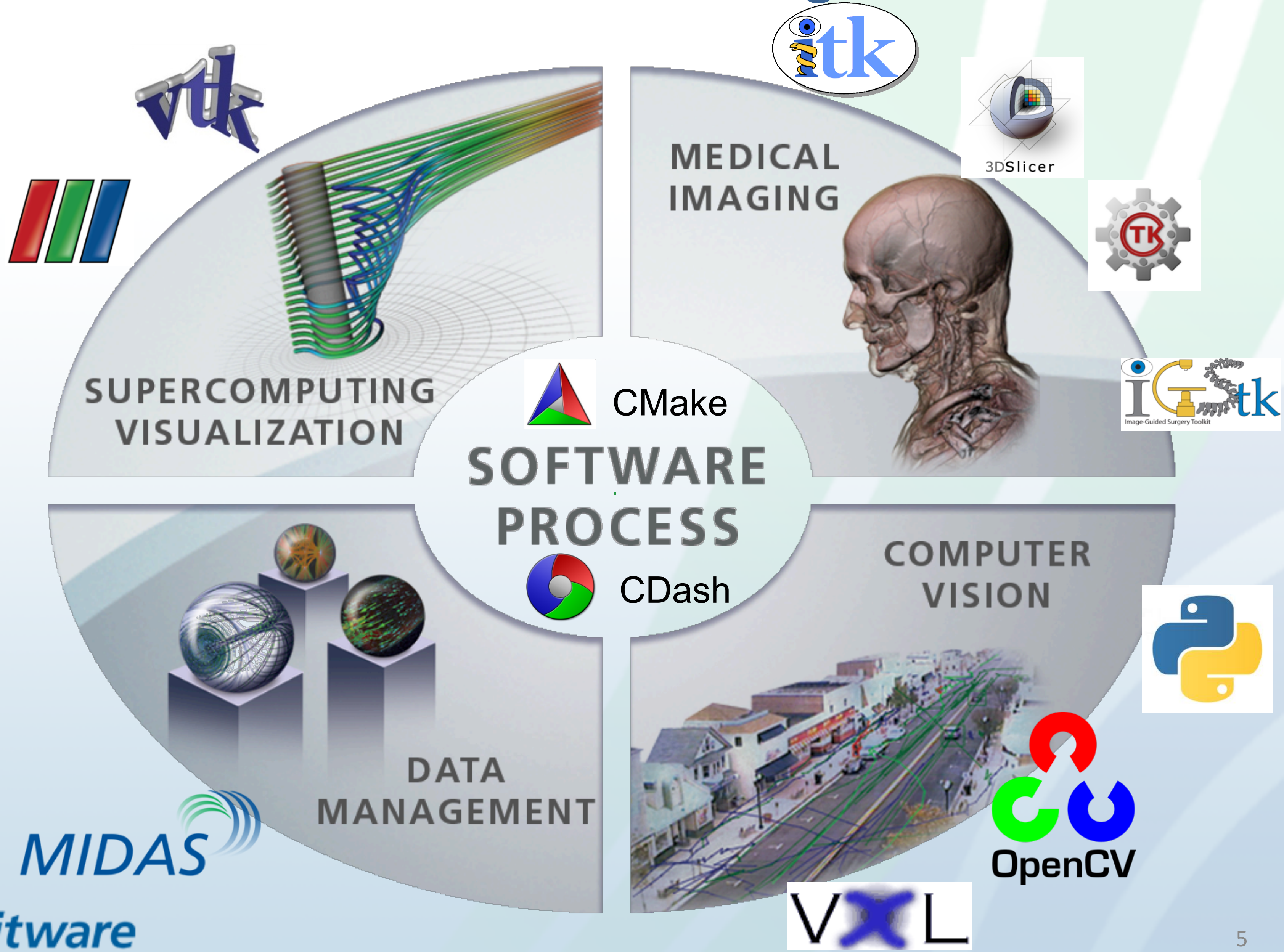
Kitware

- Founded in 1998: 5 former GE Research employees
- 95 employees: 42% PhD
- Privately held, profitable from creation, no debt
- Rapidly Growing: >30% in 2010, 7M web-visitors/quarter
- Offices
  - Albany, NY
  - **Carrboro, NC**
  - Lyon, France
  - Bangalore, India

2006 National Tibbetts Award
Presented to
Loki Incorporated
In Recognition of your Outstanding Contributions to the
SBIR Program
SBTC

- 2011 Small Business Administration’s Tibbetts Award
- HPCWire Readers and Editor’s Choice
- Inc’s 5000 List: 2008 to 2010

Kitware: Core Technologies



SUPERCOMPUTING VISUALIZATION

MEDICAL IMAGING

SOFTWARE PROCESS

CMake

CDash

DATA MANAGEMENT

COMPUTER VISION

Opening Up Chemistry

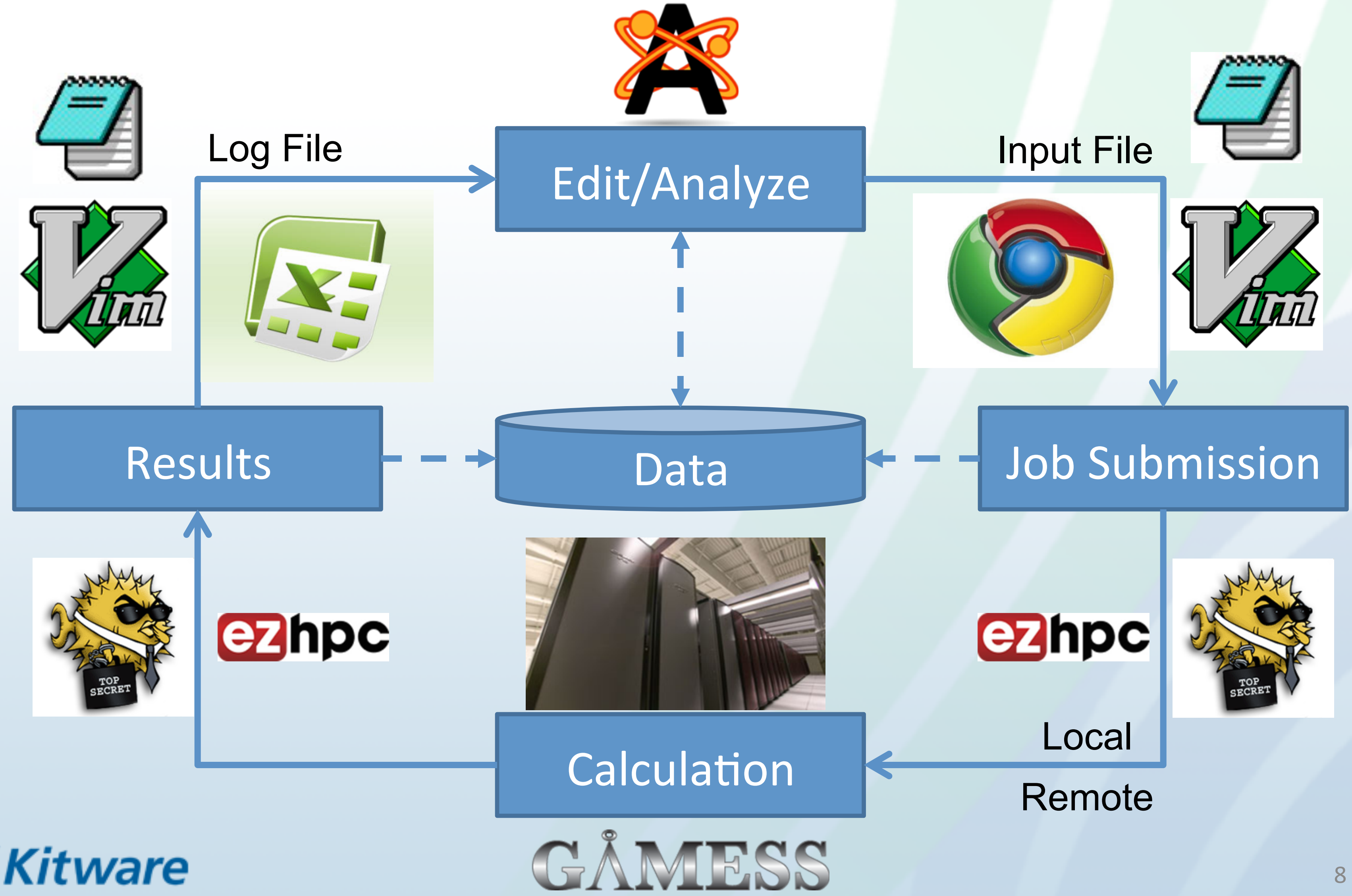
- Computational chemistry is currently one of the more closed sciences
- Lots of black box proprietary codes
  - Only a few have access to the code
  - Publishing results from black box codes
  - Many file formats in use, little agreement
- More papers should be including data
- Growing need for open standards

Movements for Open Chemistry

- Formed an “unorganization” – Blue Obelisk
  - Published first article in 2005
  - Open data, open standards and open source
  - Meet at ACS and other conferences when possible
  - Follow-up article currently in press

- Quixote collaboration more recently
  - Provide meaningful data storage and exchange
  - Principally targeting computational chemistry

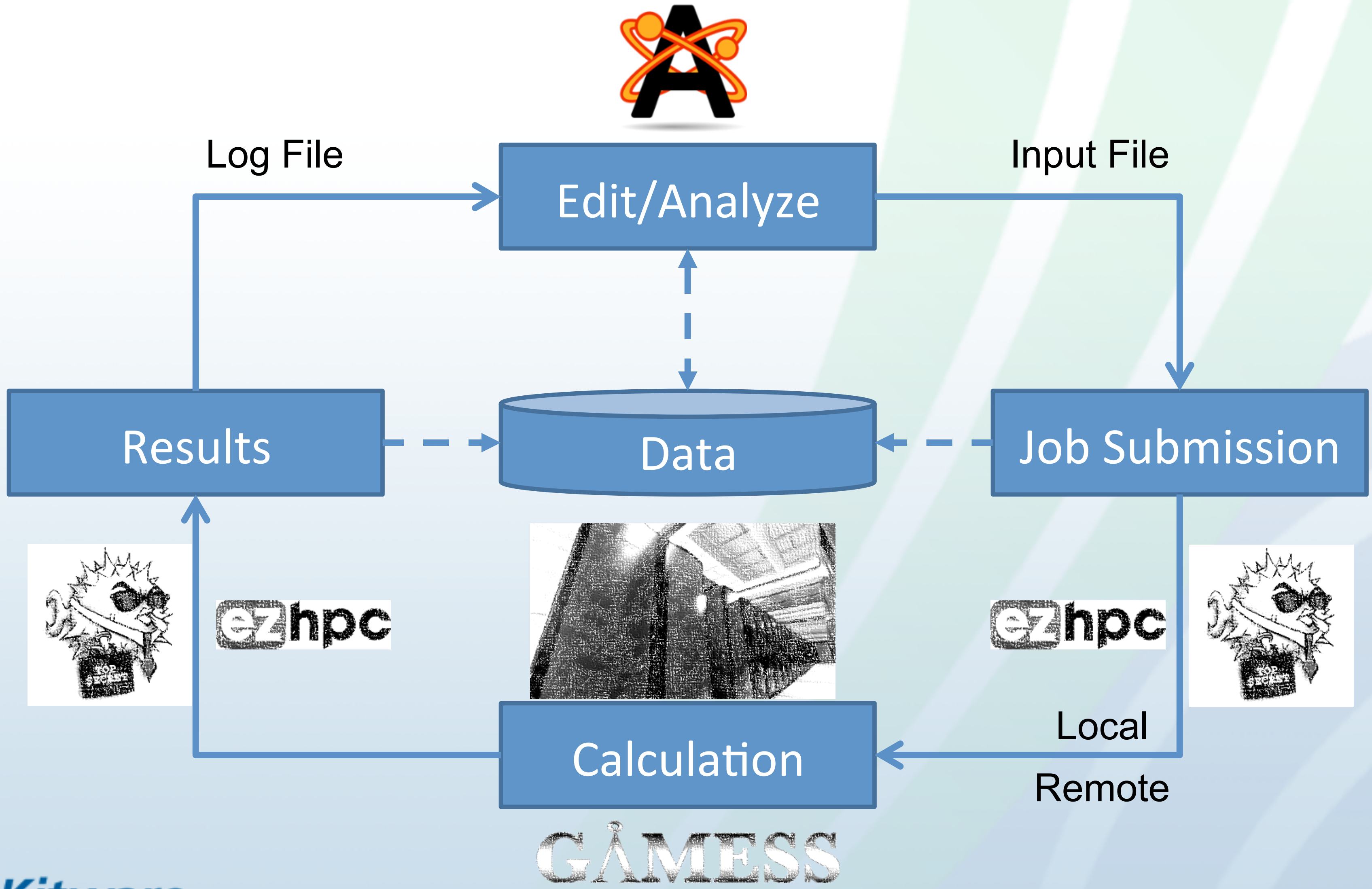
Typical Chemistry Workflow



Problem: Pretty Complex/Manual

- Most steps require user intervention
- Obtain starting structure (previous work, databases)
- Edit structure
- Write input file
- Move input file to cluster
- Submit to queue
- Wait for completion
- Retrieve input file
- Analyze output file
- Extract the relevant data, change formats
- Store results
- Repeat

Improved Chemistry Workflow



Log File

Edit/Analyze

Input File

Results

Data

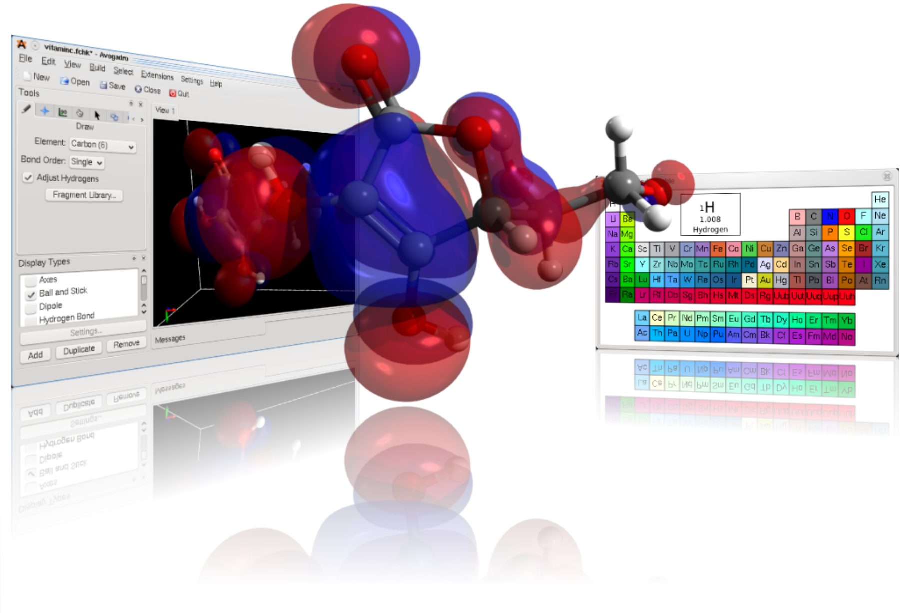
Job Submission

Calculation

Local

Remote

Avogadro



Element: Carbon (6)

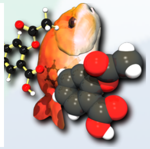
Bond Order: Single

☑ Adjust Hydrogens

- Project began 2006
- Split into library and application (plugin based)
- One of very few open source **editors**
- Designed to be extensible from the start
- Generate input & read output from many codes
- An active and growing community
- Chemistry needs a free, open framework

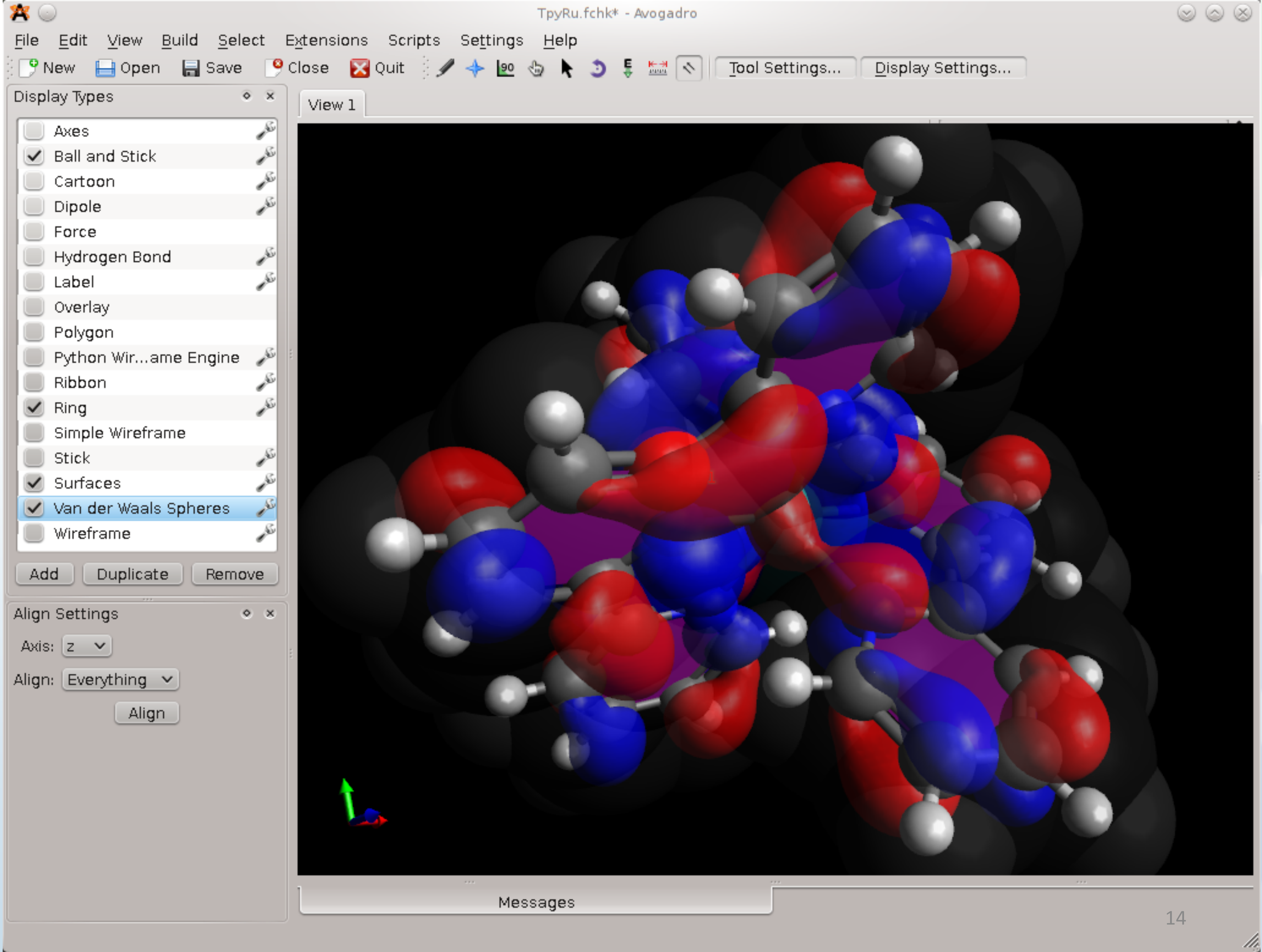
Avogadro's Roots

- Avogadro projected started in 2006
- First funded work in 2007 by Marcus Hanwell
    - Google Summer of Code student
    - Final year of Ph.D. spent the summer coding
    - Funded as part of KDE project – Kalzium editor
- Built on several other open source projects
    - Qt, Eigen, Open Babel, Blue Obelisk Data Repository
- Also uses open standards, e.g. OpenGL
- Cross platform, open source stack



Avogadro Vital Statistics

- Supports Linux, Windows and Mac OS X
- Contributions from over 20 developers
- Over 180,000 downloads over 4 years
- Translated into 19 languages
- Used by Kalzium for molecular editor
- Featured by Trolltech/Nokia,
  - Qt in use
  - Qt ambassador program



Display Types

- ☐ Axes
- ☑ Ball and Stick
- ☐ Cartoon
- ☐ Dipole
- ☐ Force
- ☐ Hydrogen Bond
- ☐ Label
- ☐ Overlay
- ☐ Polygon
- ☐ Python Wir...ame Engine
- ☐ Ribbon
- ☑ Ring
- ☐ Simple Wireframe
- ☐ Stick
- ☑ Surfaces
- ☑ Van der Waals Spheres
- ☐ Wireframe

Align Settings

Axis: z

Align: Everything

Desktop Database

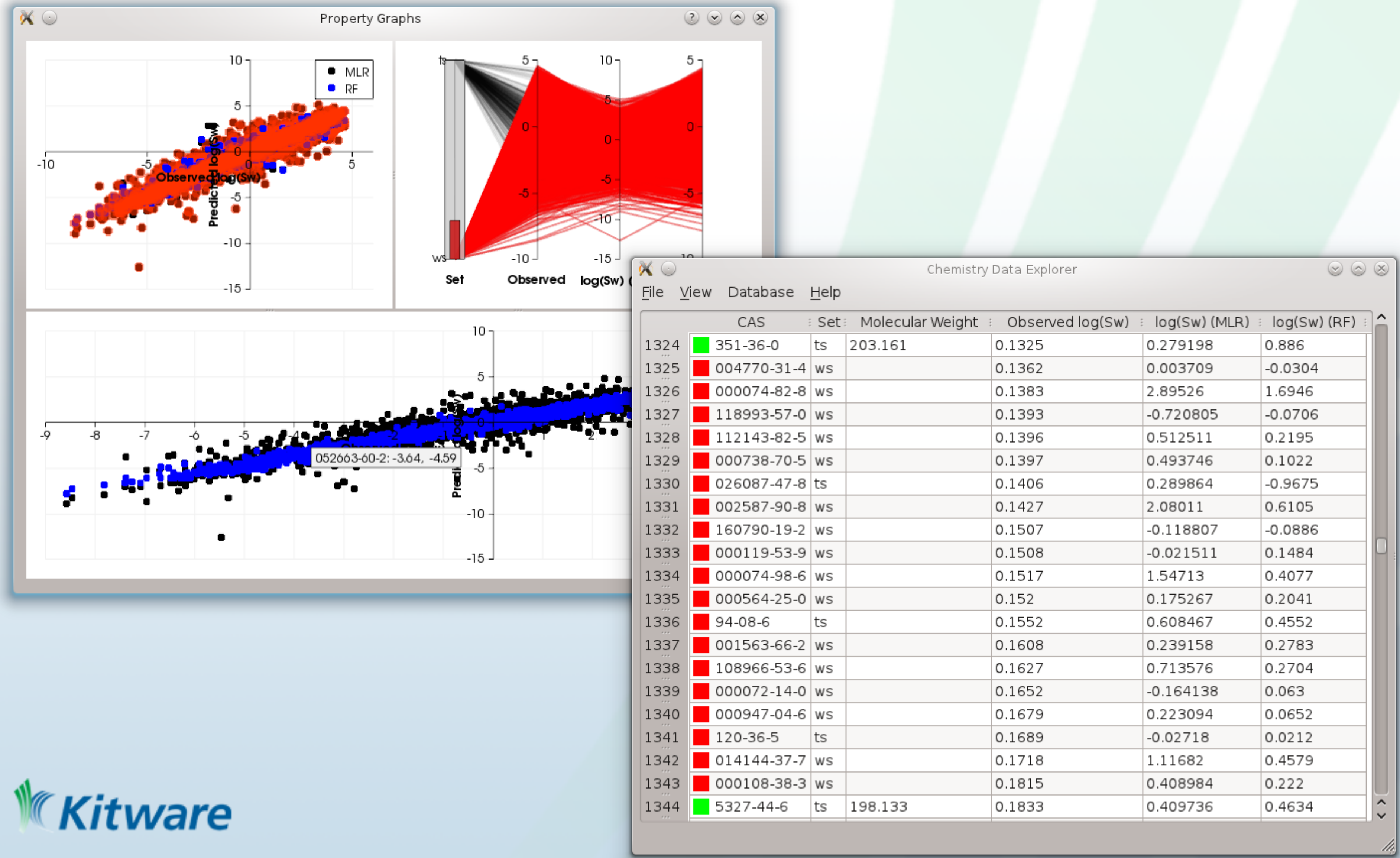
mongoDB

- Use of "document store" NoSQL
  - Doesn't force too much structure
    - Some entries have experimental data available
    - Some have computational jobs
  - Employ a "pile of stuff" approach
    - Can store both source and derived data
    - Calculate identifiers, QSAR properties, etc
- MongoDB is a scalable, open solution
  - Proven scaling with large web applications

Chemistry Data Explorer

- Qt application
- Connects to local or remote database
- Uses VTK for visual data exploration
- Can ingest new data
  - Uses Open Babel to generate descriptors
  - Standard InChi, SMILES, molecular weight
  - More could be added
    - All derived from files stored in the database

Chemistry Data Explorer

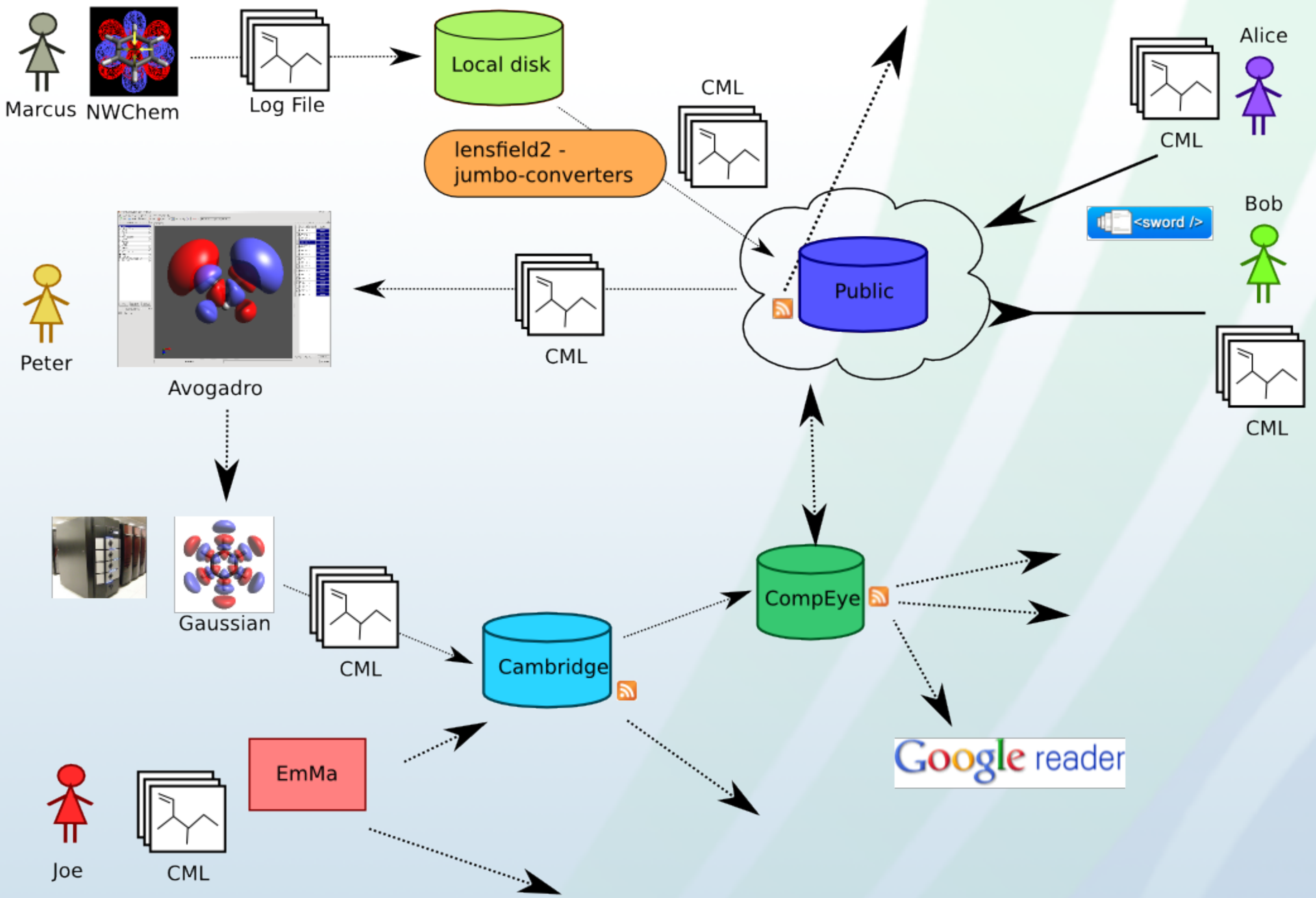


	CAS	Set	Molecular Weight	Observed log(Sw)	log(Sw) (MLR)	log(Sw) (RF)
1324	🟩 351-36-0	ts	203.161	0.1325	0.279198	0.886
1325	🟥 004770-31-4	ws		0.1362	0.003709	-0.0304
1326	🟥 000074-82-8	ws		0.1383	2.89526	1.6946
1327	🟥 118993-57-0	ws		0.1393	-0.720805	-0.0706
1328	🟥 112143-82-5	ws		0.1396	0.512511	0.2195
1329	🟥 000738-70-5	ws		0.1397	0.493746	0.1022
1330	🟥 026087-47-8	ts		0.1406	0.289864	-0.9675
1331	🟥 002587-90-8	ws		0.1427	2.08011	0.6105
1332	🟥 160790-19-2	ws		0.1507	-0.118807	-0.0886
1333	🟥 000119-53-9	ws		0.1508	-0.021511	0.1484
1334	🟥 000074-98-6	ws		0.1517	1.54713	0.4077
1335	🟥 000564-25-0	ws		0.152	0.175267	0.2041
1336	🟥 94-08-6	ts		0.1552	0.608467	0.4552
1337	🟥 001563-66-2	ws		0.1608	0.239158	0.2783
1338	🟥 108966-53-6	ws		0.1627	0.713576	0.2704
1339	🟥 000072-14-0	ws		0.1652	-0.164138	0.063
1340	🟥 000947-04-6	ws		0.1679	0.223094	0.0652
1341	🟥 120-36-5	ts		0.1689	-0.02718	0.0212
1342	🟥 014144-37-7	ws		0.1718	1.11682	0.4579
1343	🟥 000108-38-3	ws		0.1815	0.408984	0.222
1344	🟩 5327-44-6	ts	198.133	0.1833	0.409736	0.4634

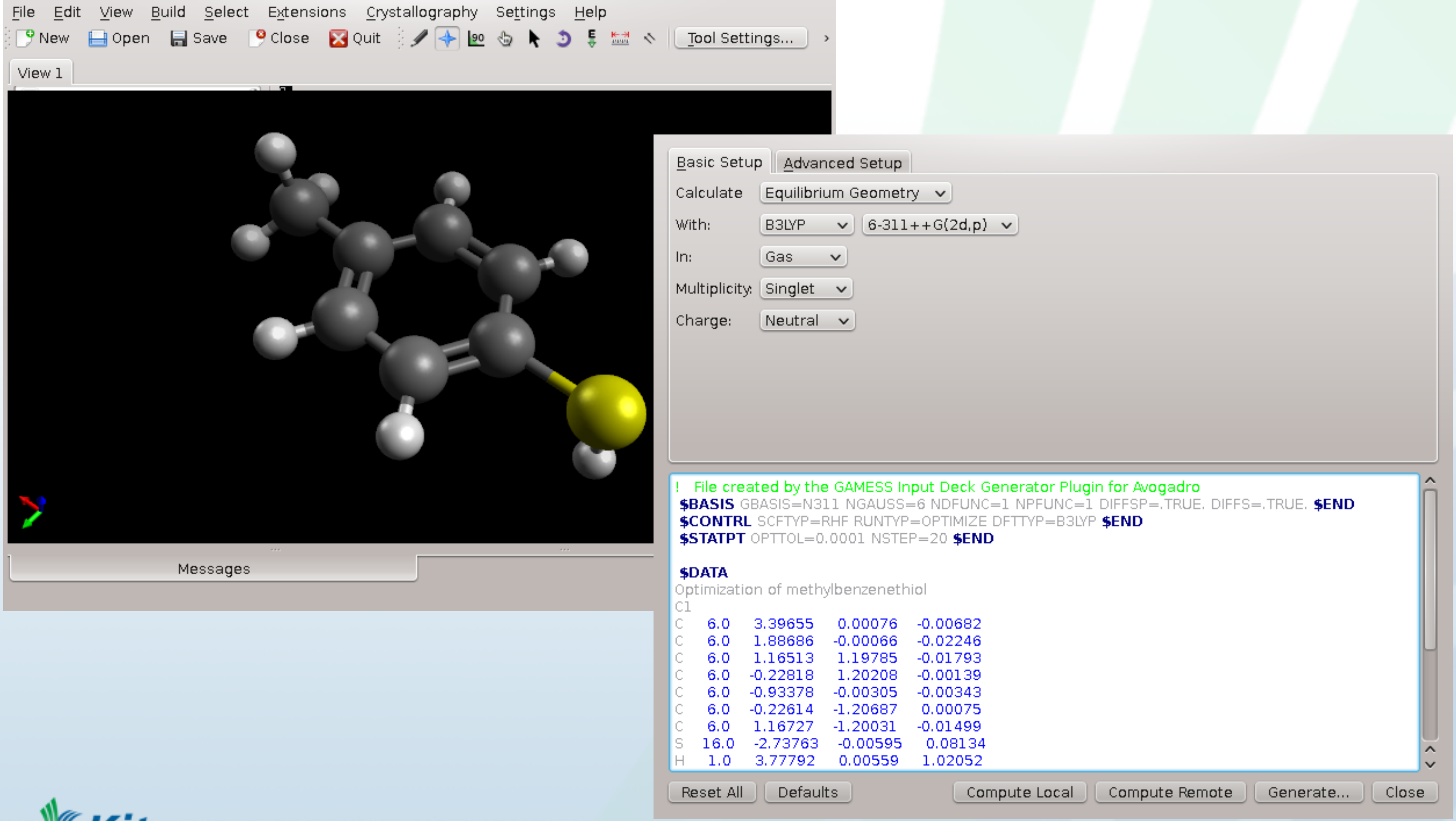
Database Interaction on the Web

- Avogadro directly accesses some (read-only) public databases:
  - PDB, NIH “fetch by name”
  - Resolve structure to common name using CIR
  - More could be added
- ChemData also uses NIH CIR for data
- Quixote aims to support both public and private sharing models – open framework

Quixote Architecture



Avogadro → GAMESS



Calculate: Equilibrium Geometry

With: B3LYP 6-311++G(2d,p)

In: Gas

Multiplicity: Singlet

Charge: Neutral

```
!   File created by the GAMESS Input Deck Generator Plugin for Avogadro
 $BASIS GBASIS=N311 NGAUSS=6 NDFUNC=1 NPFUNC=1 DIFFSP=.TRUE. DIFFS=.TRUE. $END
 $CONTRL SCFTYP=RHF RUNTYP=OPTIMIZE DFTTYP=B3LYP $END
 $STATPT OPTTOL=0.0001 NSTEP=20 $END

 $DATA
Optimization of methylbenzenethiol
C1
C     6.0     3.39655     0.00076    -0.00682
C     6.0     1.88686    -0.00066    -0.02246
C     6.0     1.16513     1.19785    -0.01793
C     6.0    -0.22818     1.20208    -0.00139
C     6.0    -0.93378    -0.00305    -0.00343
C     6.0    -0.22614    -1.20687     0.00075
C     6.0     1.16727    -1.20031    -0.01499
S    16.0    -2.73763    -0.00595     0.08134
H     1.0     3.77792     0.00559     1.02052
```

OpenQube – Quantum Data

- Reads in key quantum data
  - Basis set used in calculation
  - Eigenvectors for molecular orbitals
  - Density matrix for electron density
  - Standard geometry
- Multithreaded calculation
  - Produce regular grids of scalar data
  - Molecular orbitals, electron density…

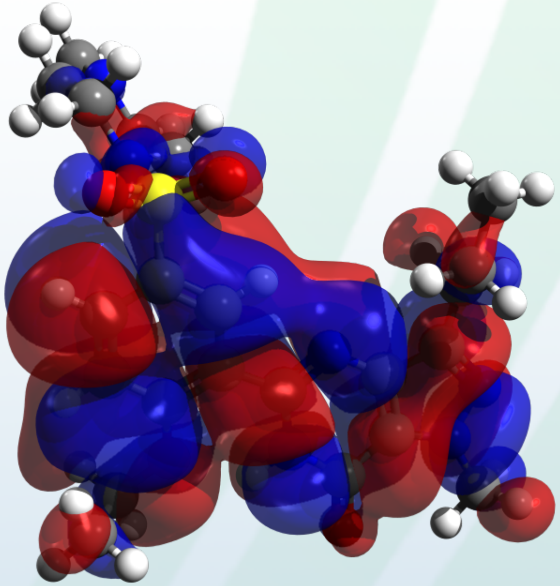
Molecular Orbitals and Electron Density

- Quantum files store basis sets and matrices

$$GTO = ce^{-\alpha r^2}$$

$$\phi_i = \sum_{\mu} c_{\mu i} \phi_{\mu}$$

$$\rho(r) = \sum_{\mu} \sum_{\nu} P_{\mu\nu} \phi_{\mu} \phi_{\nu}$$



- Using these equations, and the supplied matrices – calculate cubes

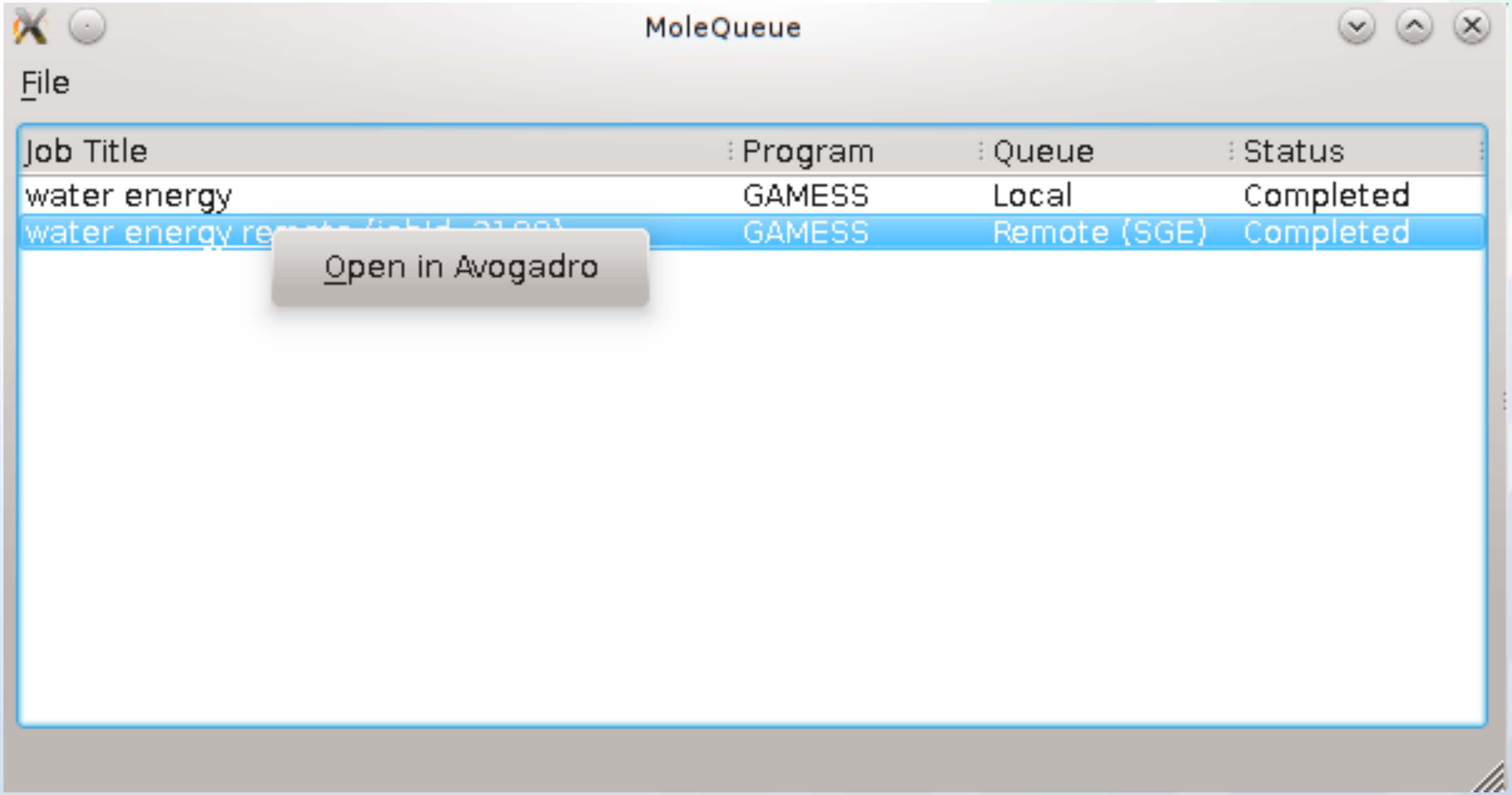
Calling Stand Alone Programs

- Many already supported:
  - GAMESS, GAMESS-UK, Molpro, Q-Chem, MOPAC, NWChem, Gaussian, Dalton
  - Easy to add more
- Some codes writing Avogadro based custom applications,
  - Q-Chem, Molpro…
- DLPOLY author approached me:
  - Open sourced DLPOLY2, want a GUI

Job Submission & Management

- Take input file, submit to queue, monitor, retrieve, repeat
- System tray resident Qt application
  - Manage both local and remote jobs
- Interest from developers
  - Use in other applications
  - Share development/maintenance burden

Open in Avogadro When Complete

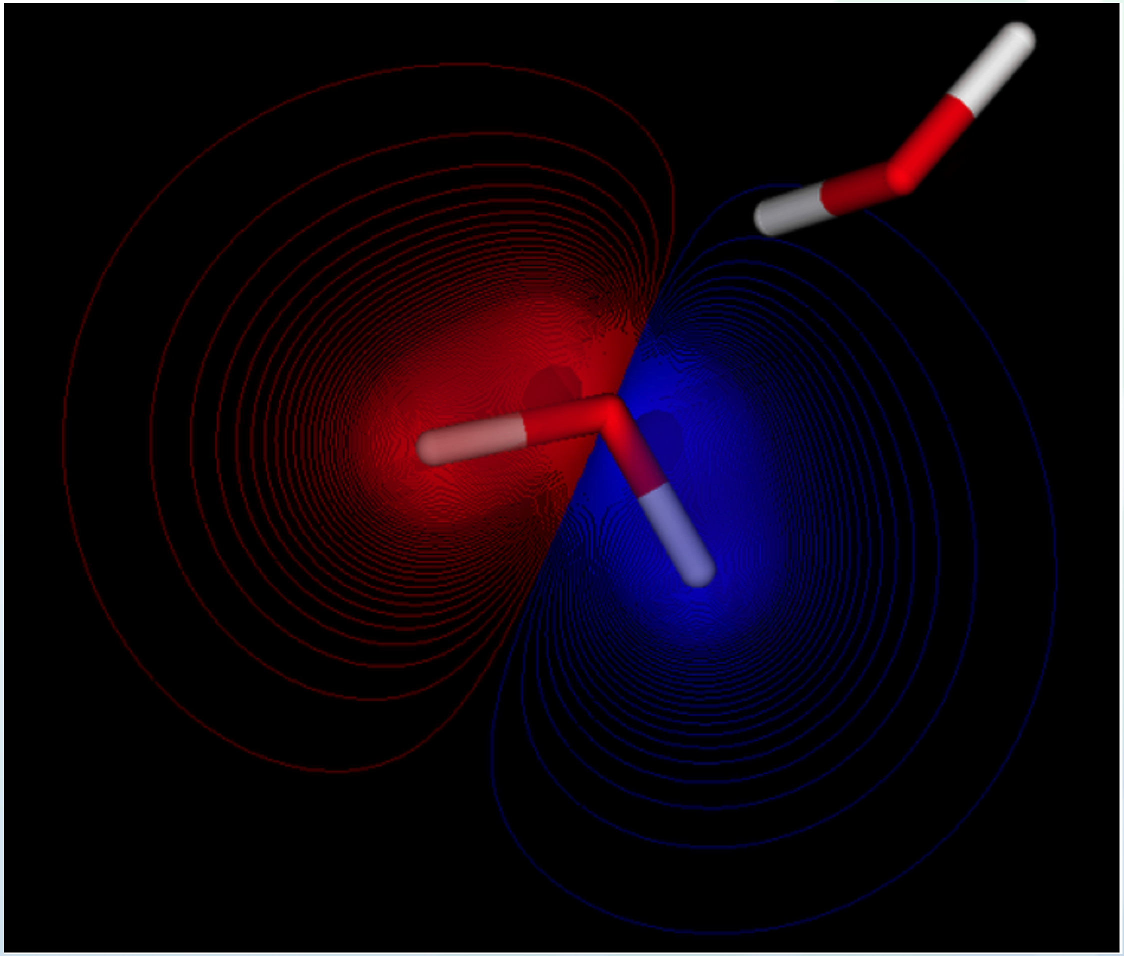


Job Title	Program	Queue	Status
water energy	GAMESS	Local	Completed
water energy re	GAMESS	Remote (SGE)	Completed

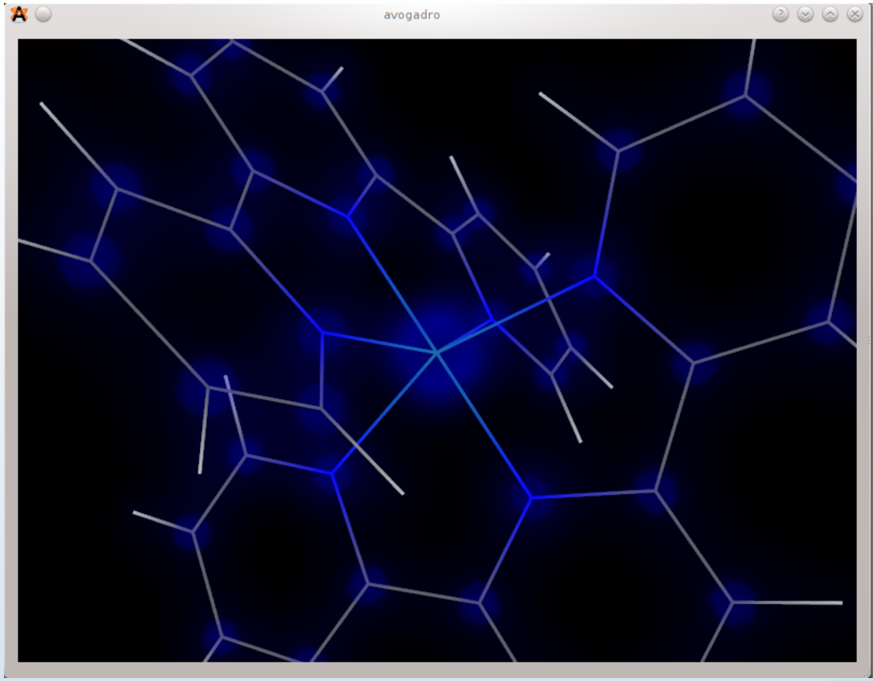
Advanced Visualization: VTK

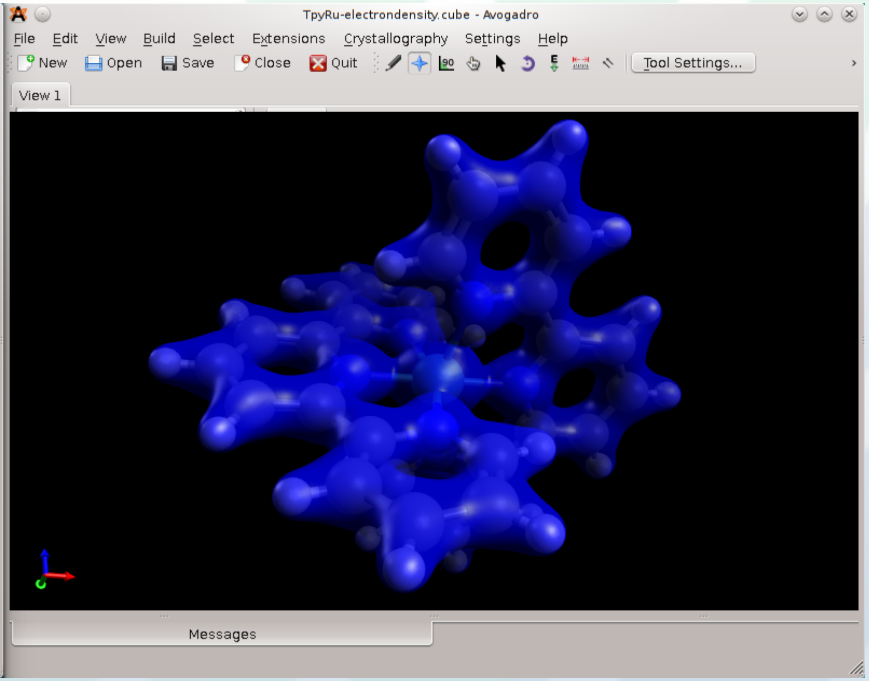
- New Avogadro plugin:
  - Takes volumetric data from Avogadro
  - Uses GPU accelerated rendering in VTK
- Excitement from many in the community
- Several groups interested in collaborating
- Google Summer of Code project
- Leverage significant capabilities in VTK

Volume Rendered With Contours

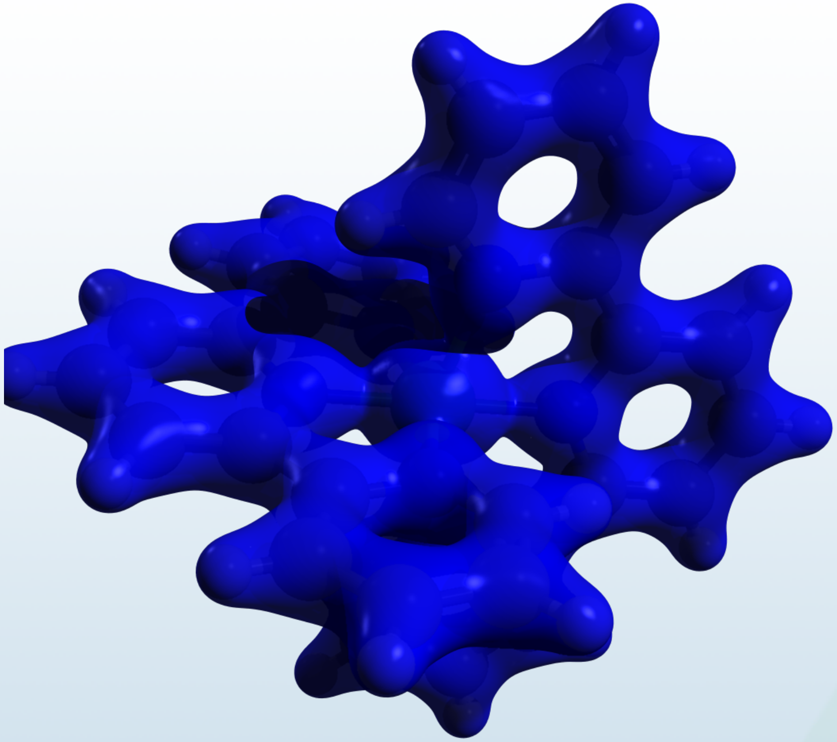


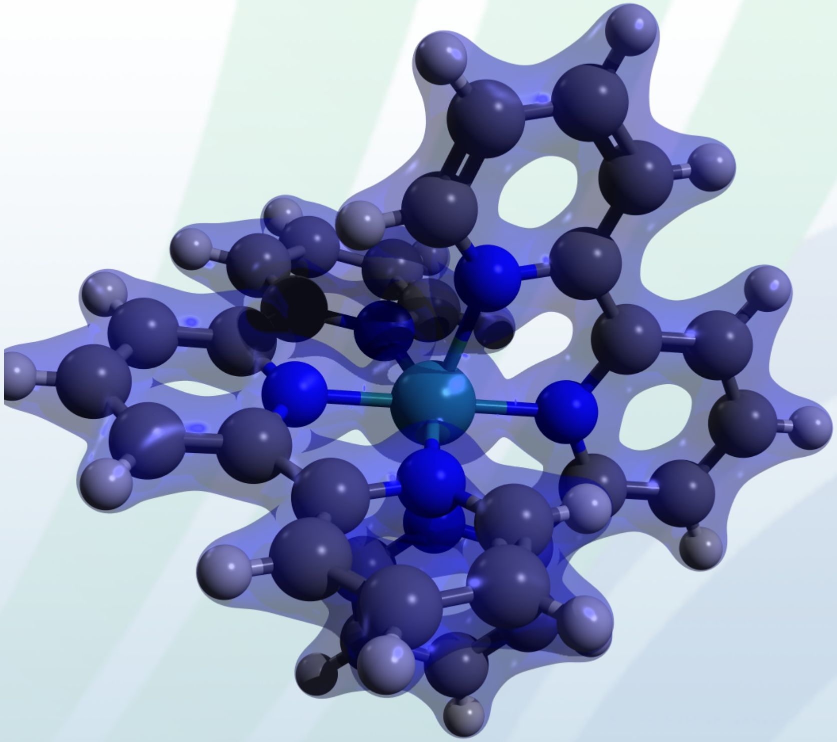
Electron Density Volume Render




TpyRu-electrondensity.cube - Avogadro
File Edit View Build Select Extensions Crystallography Settings Help
New Open Save Close Quit
Tool Settings...
View 1
Messages

Electron Density Ray Tracing





Conclusions

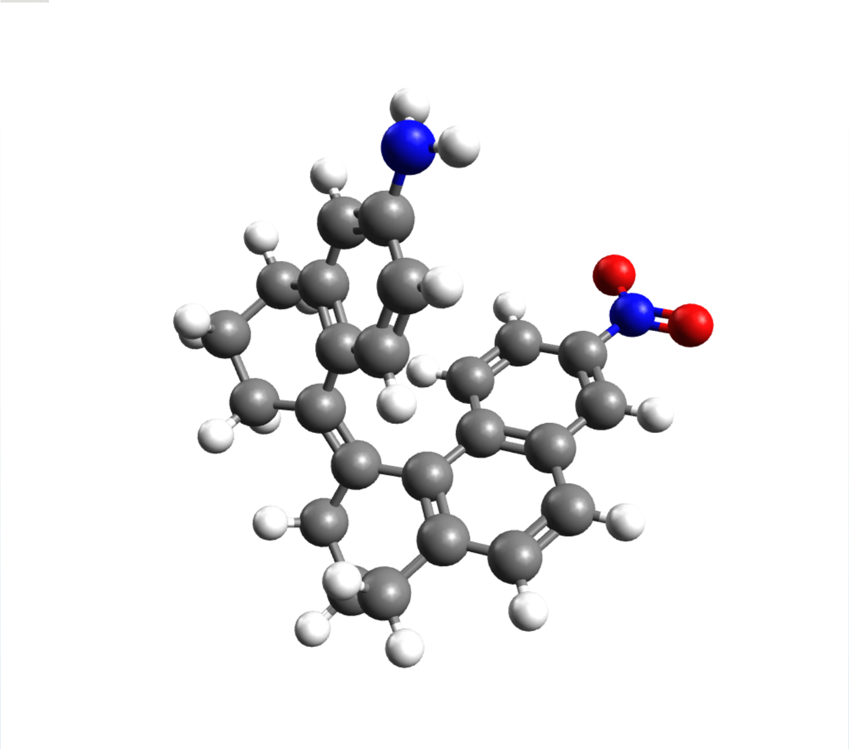
- There is still a lot of work to do
- Open databases are of critical importance
- Need tools to make retrieving and depositing data easier
- Improved data exchange is essential to improve reproducibility in chemistry
- Create shared collaboration platforms
  - Deliver improved workflows, enable research

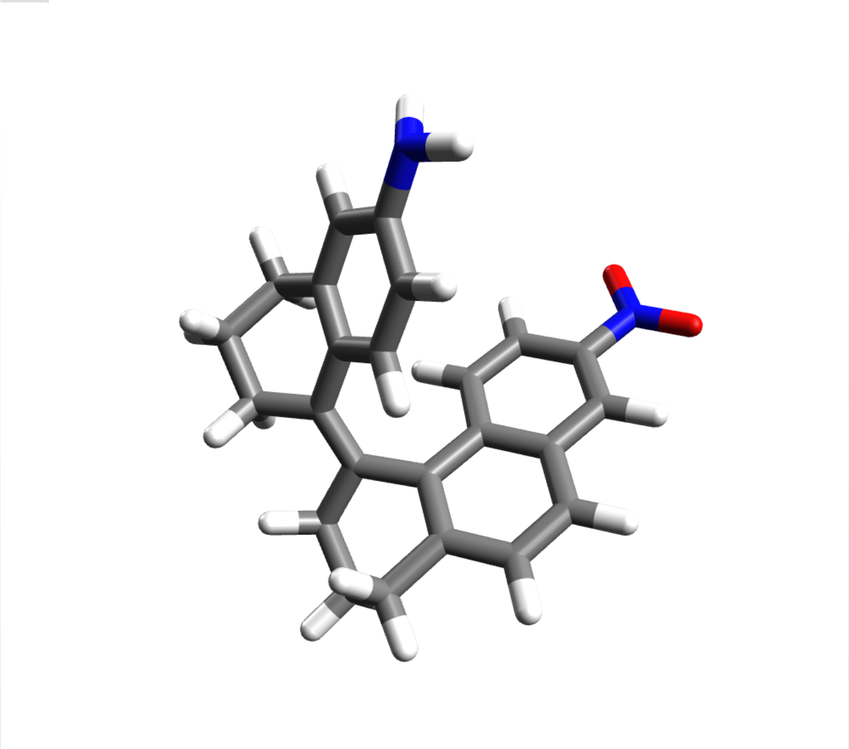
Extra Background Slides

- Additional visualization and background slides

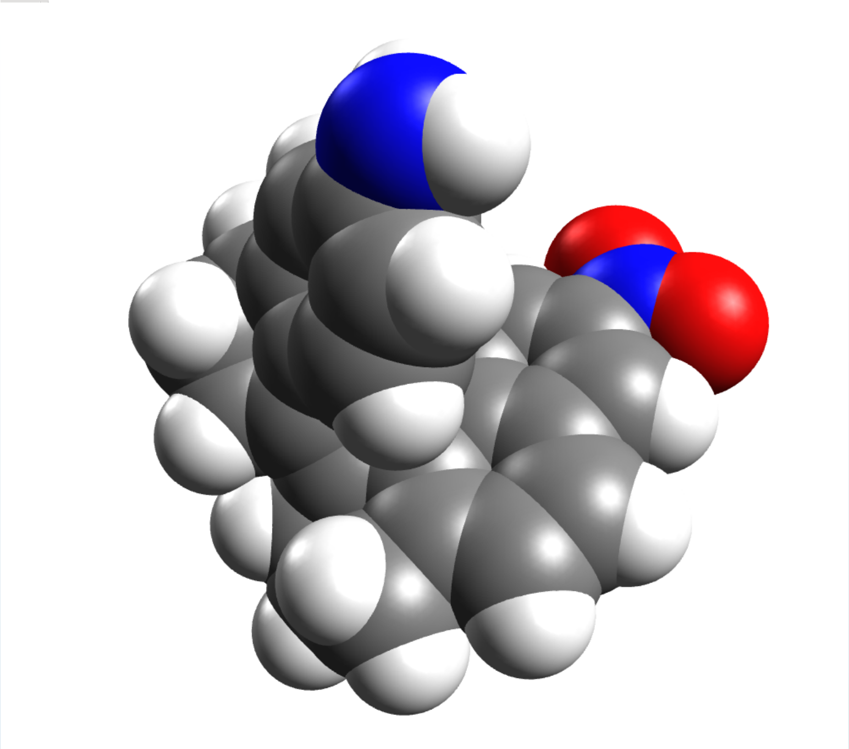
Kitware

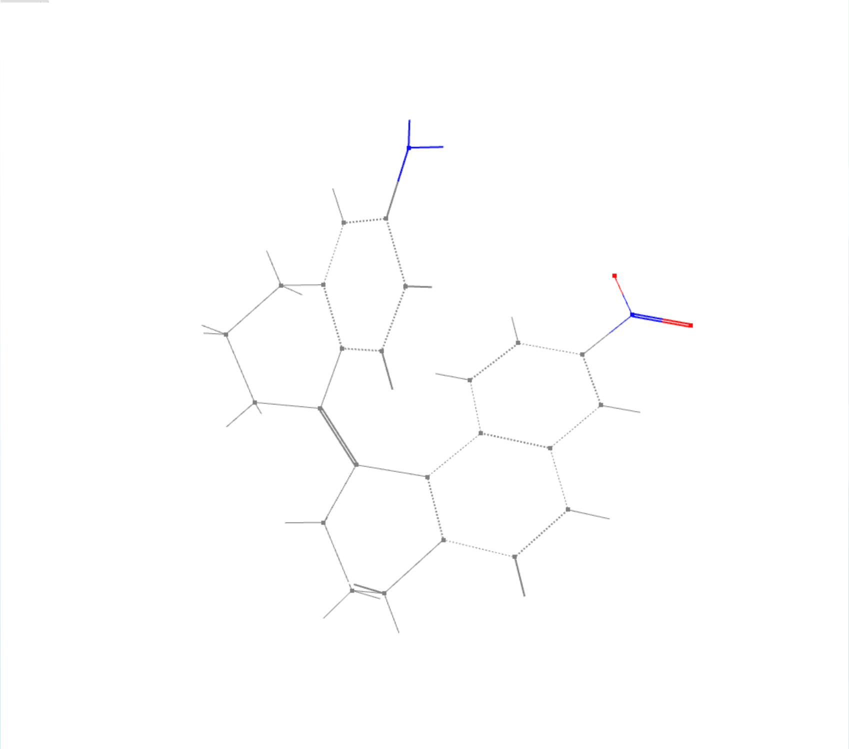
Standard Representations



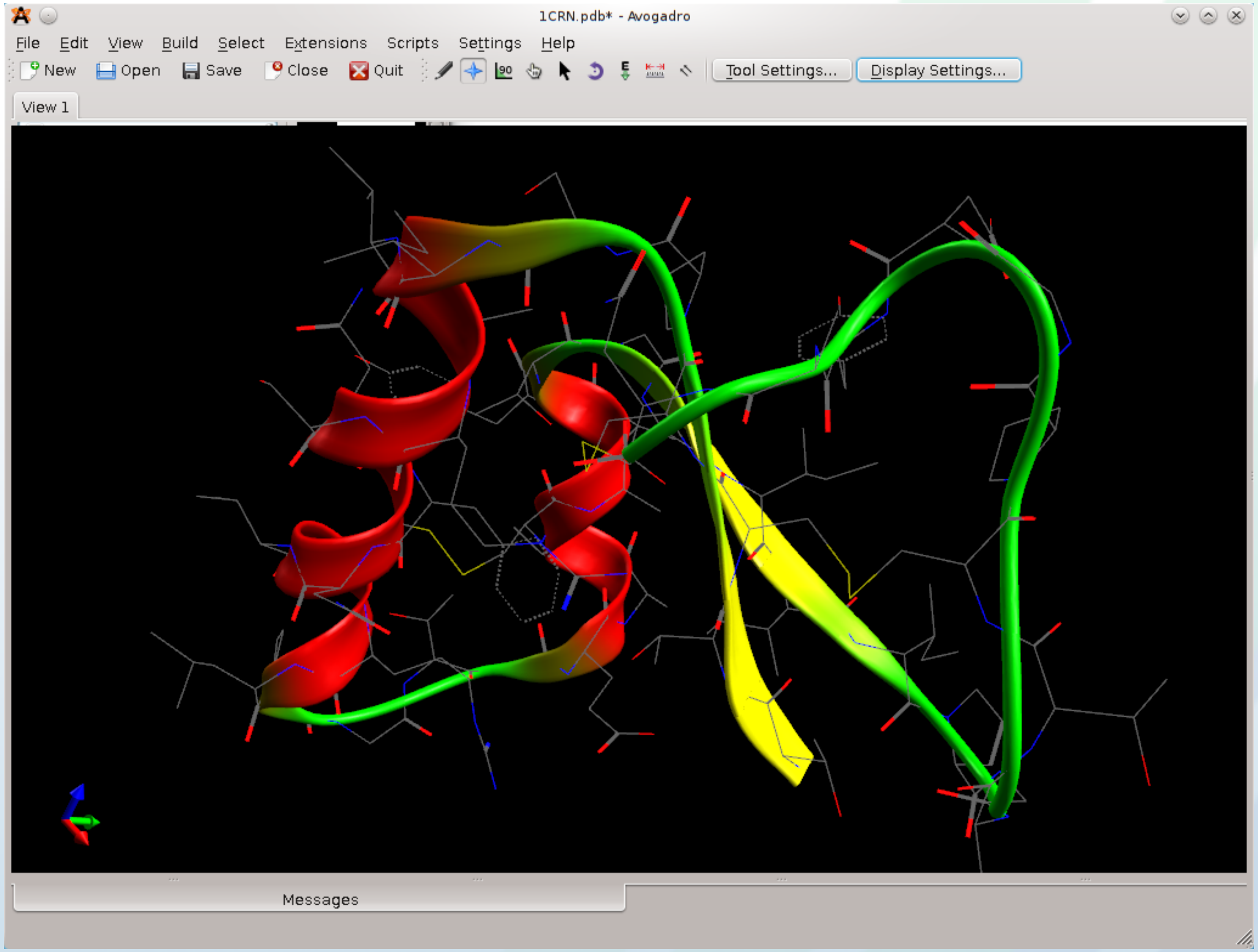


Standard Representations

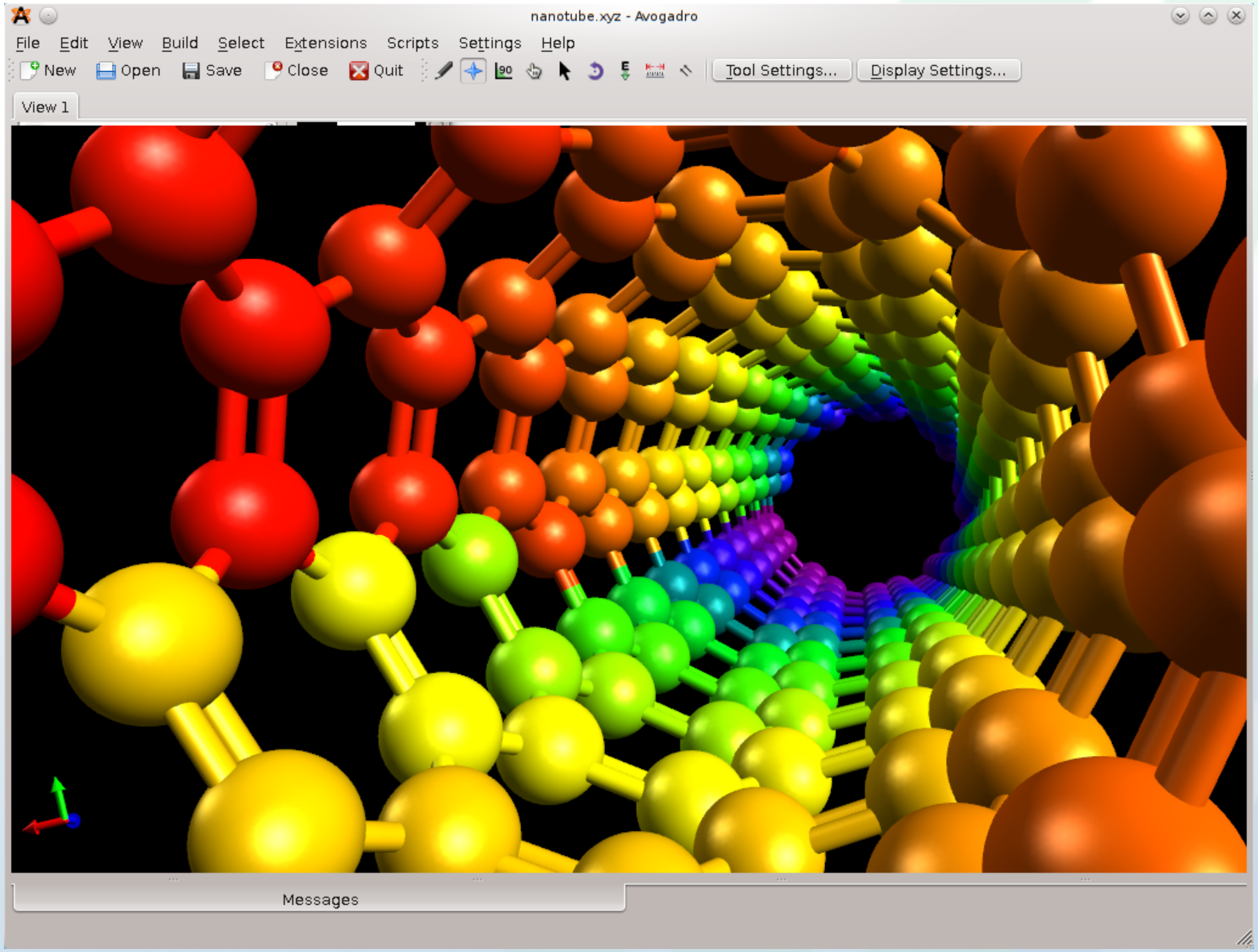




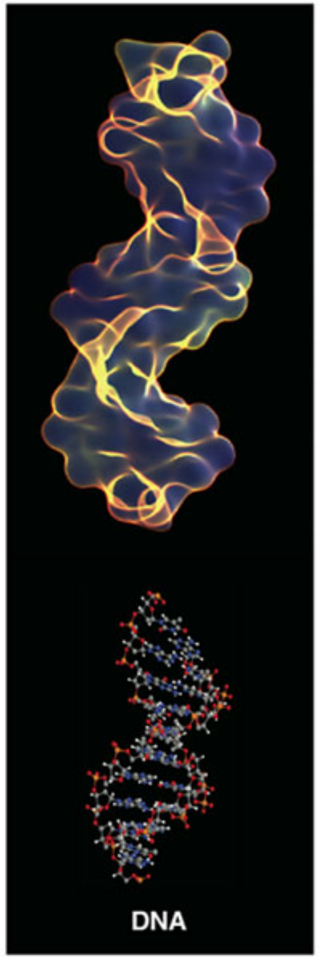
Biomolecules

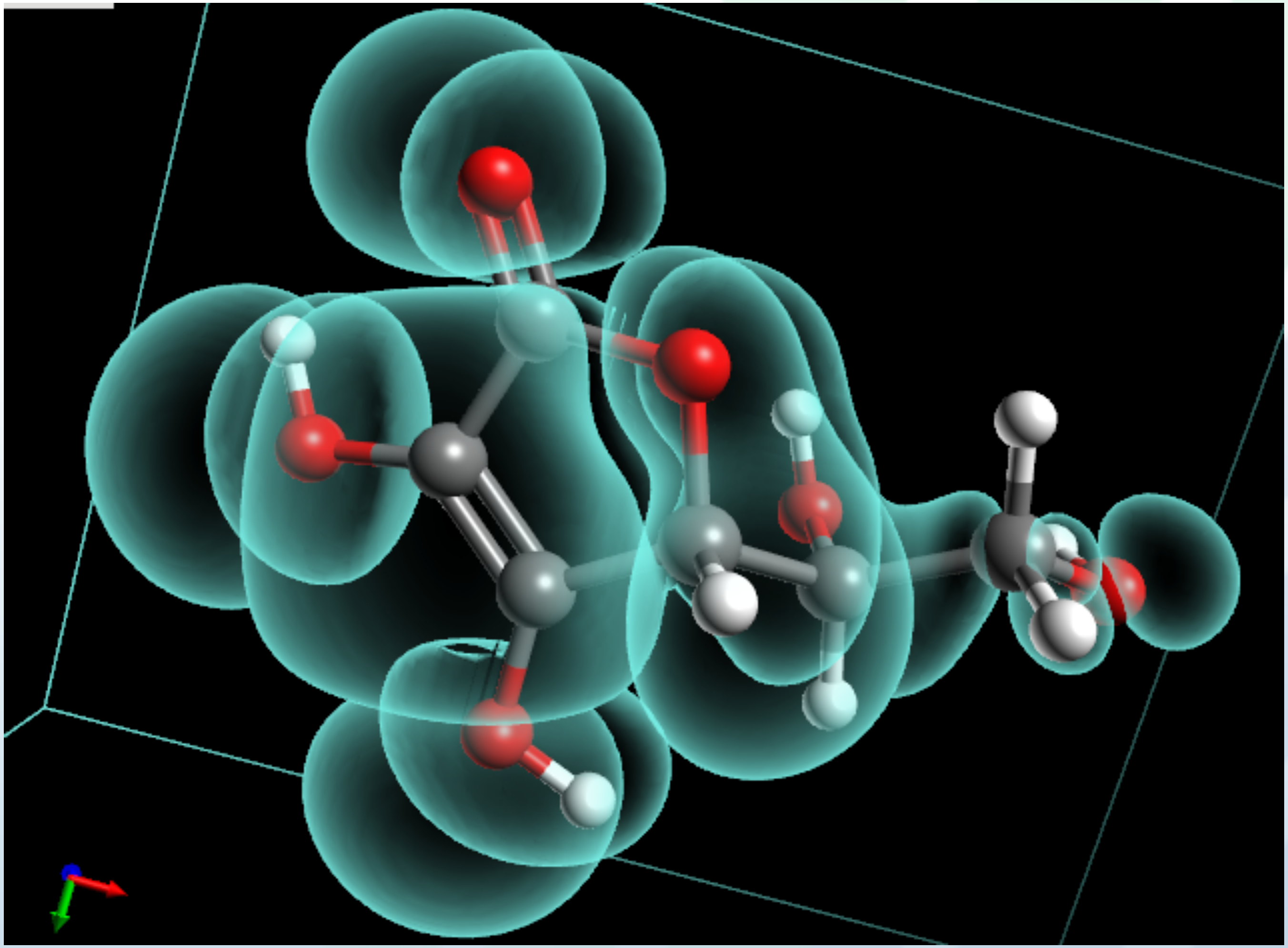

1CRN.pdb* - Avogadro
File Edit View Build Select Extensions Scripts Settings Help
New Open Save Close Quit
Tool Settings... Display Settings...
View 1
Messages

Nanomaterials

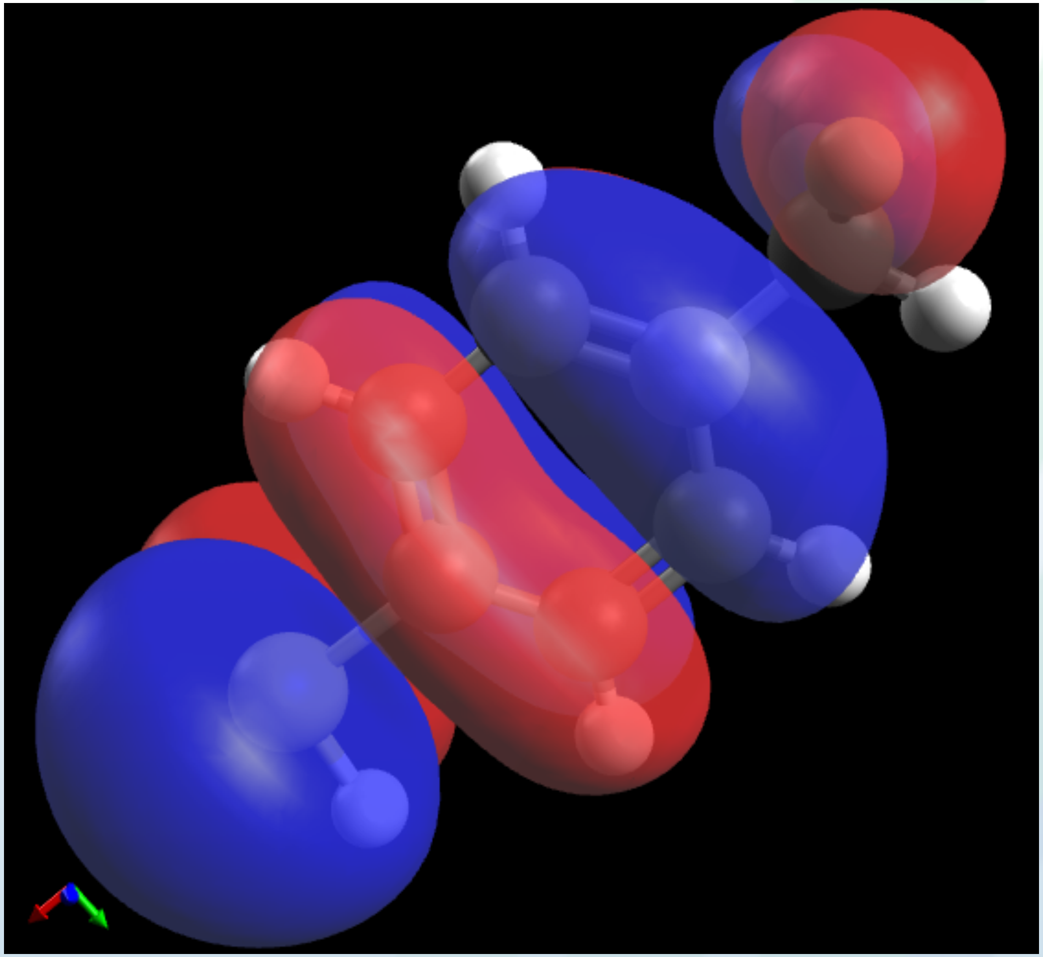


Simplified Views

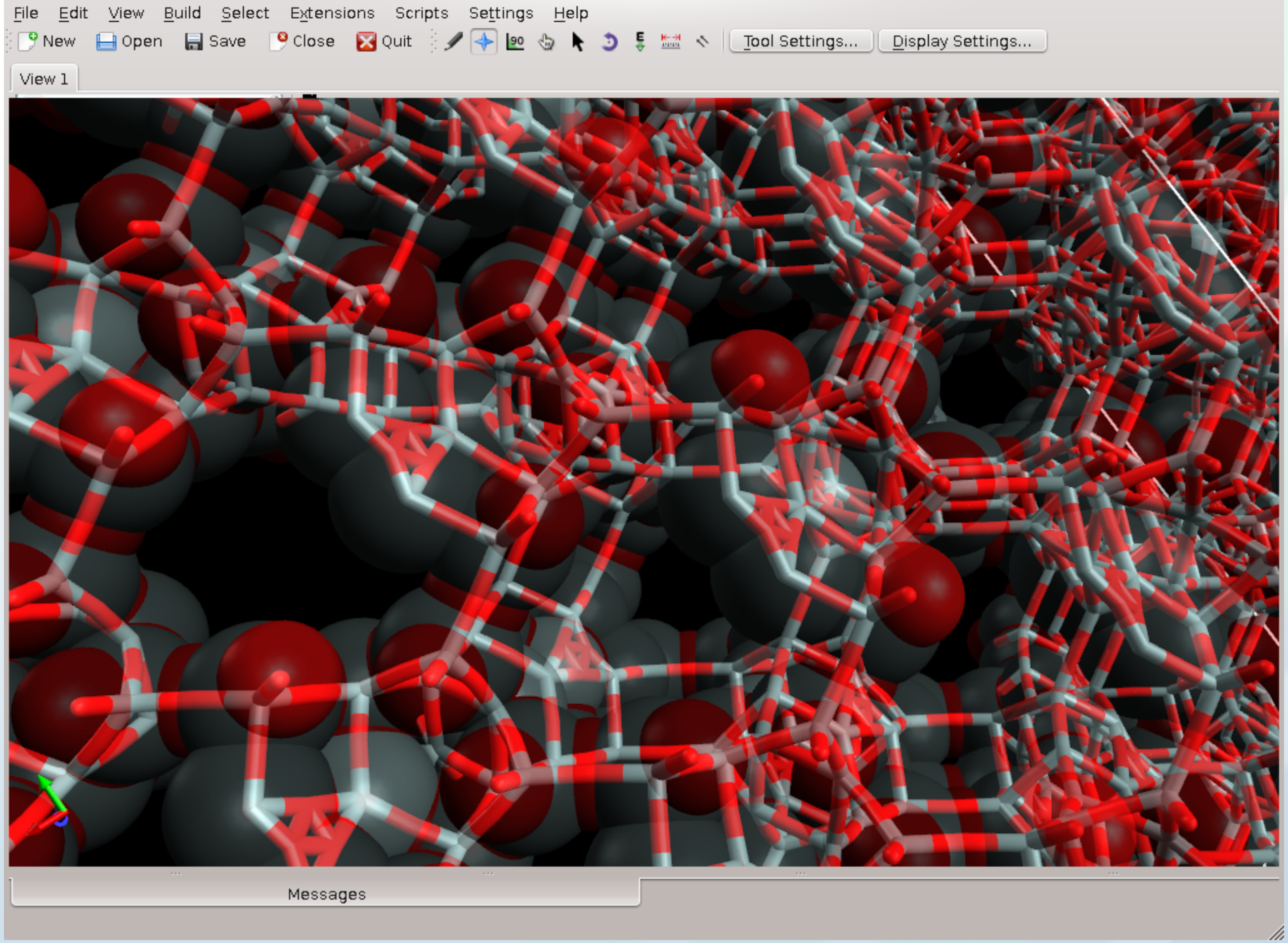




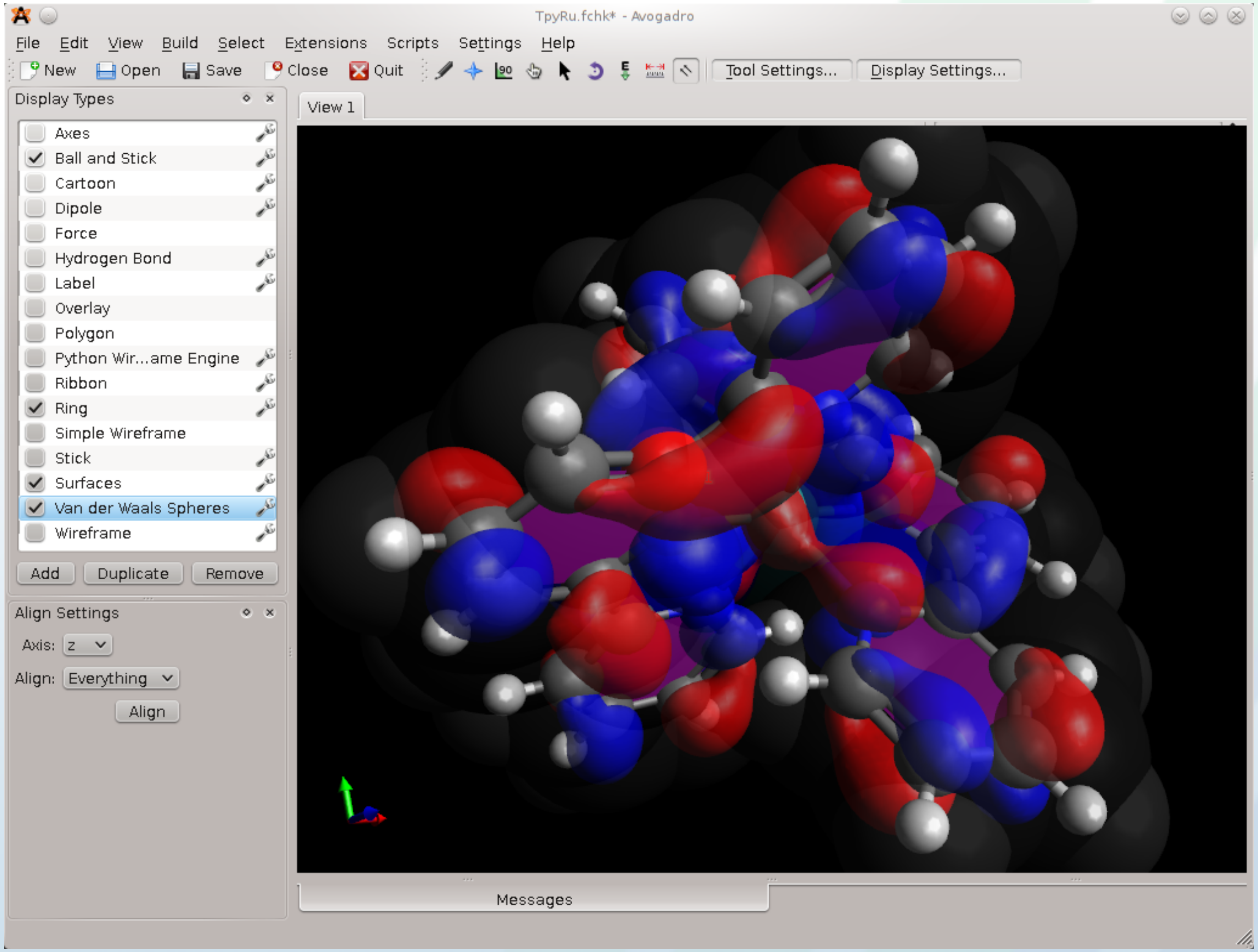
Volumetric Data: Molecular Orbitals



Periodic Systems



Hybrid Views: CPK + MO + Ball & Stick


TpyRu.fchk* - Avogadro
File Edit View Build Select Extensions Scripts Settings Help
New Open Save Close Quit
Tool Settings... Display Settings...
View 1
Add Duplicate Remove
Align
Messages

Display Types

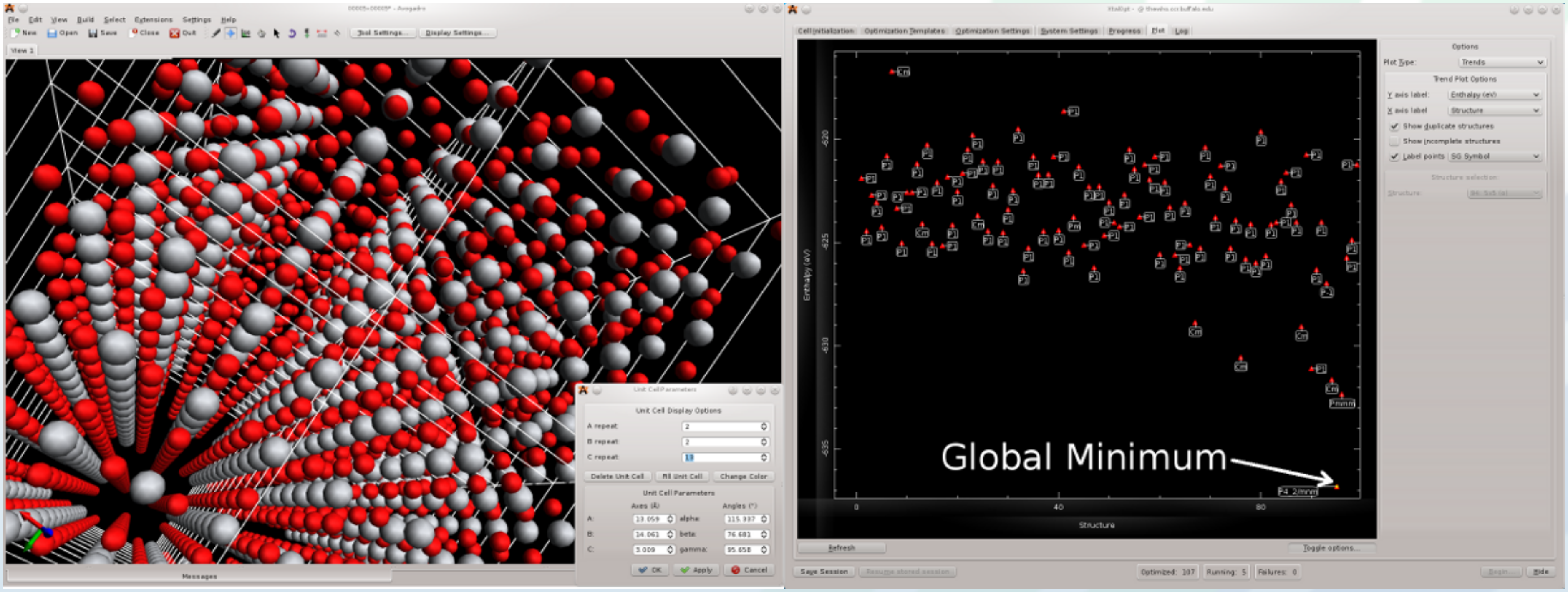
- ☐ Axes
- ☑ Ball and Stick
- ☐ Cartoon
- ☐ Dipole
- ☐ Force
- ☐ Hydrogen Bond
- ☐ Label
- ☐ Overlay
- ☐ Polygon
- ☐ Python Wir...ame Engine
- ☐ Ribbon
- ☑ Ring
- ☐ Simple Wireframe
- ☐ Stick
- ☑ Surfaces
- ☑ Van der Waals Spheres
- ☐ Wireframe

Align Settings

Axis: z

Align: Everything

Linked Views of Live Data


File Edit View Build Select Extensions Settings Help
New Open Save Close Quit
Tool Settings... Display Settings...
View 1
Messages
Cell Initialization Optimization Templates Optimization Settings System Settings Progress Plot Log
Enthalpy (eV)
-620
-625
-630
-635
0
40
80
Structure
Global Minimum
P4_2/mnm
Cm
P1
Pm
Pmmn
Refresh
Toggle options...
Save Session
Resume stored session
Begin...
Hide

Unit Cell Parameters

Unit Cell Display Options

A repeat: 2

B repeat: 2

C repeat: 13

Delete Unit Cell | Fill Unit Cell | Change Color

Unit Cell Parameters

	Axes (Å)		Angles (°)
A:	13.059	alpha:	115.337
B:	14.061	beta:	76.681
C:	3.009	gamma:	95.658

OK | Apply | Cancel

Options

Plot Type: Trends

Trend Plot Options

Y axis label: Enthalpy (eV)

X axis label: Structure

☑ Show duplicate structures

☐ Show incomplete structures

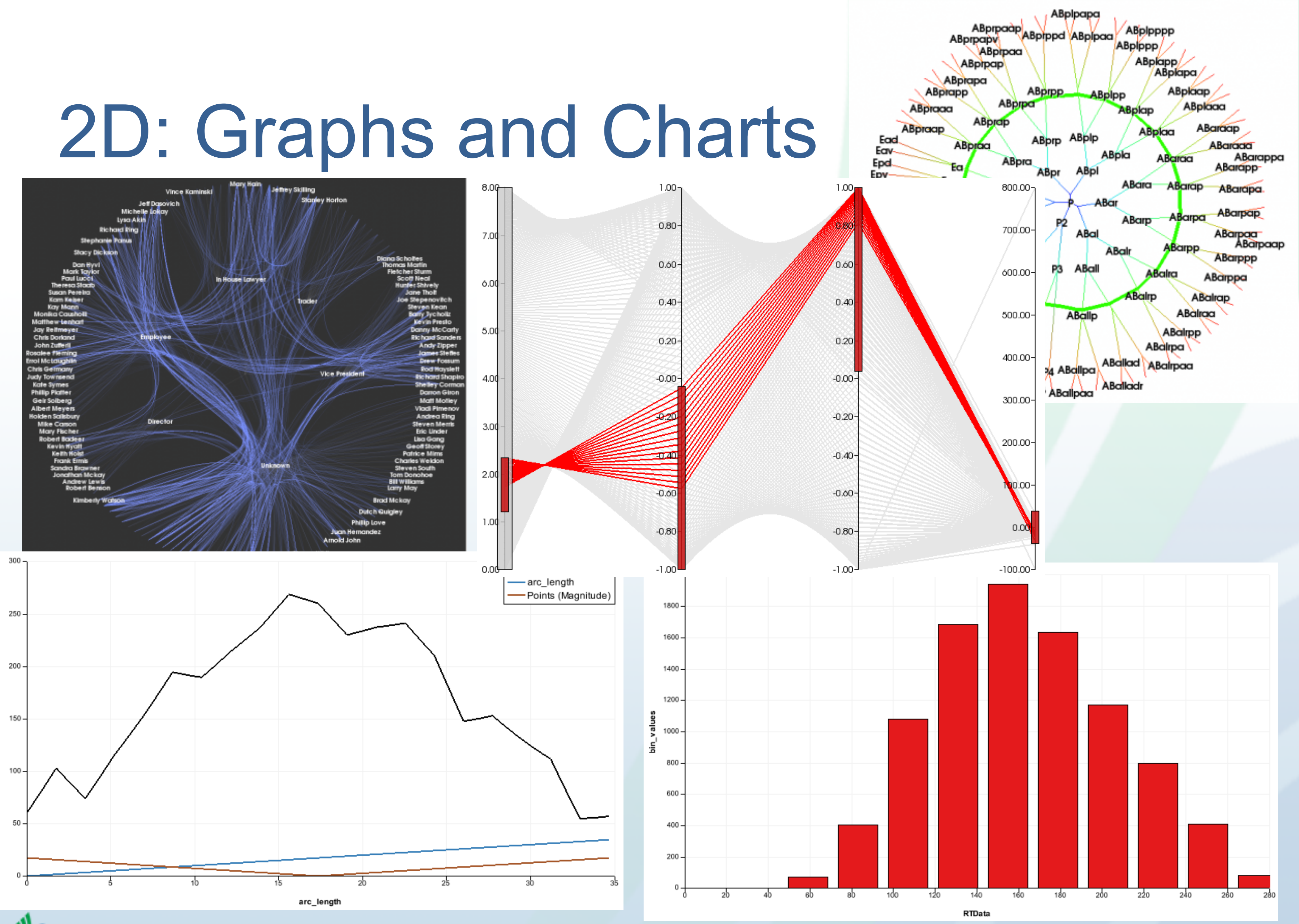
☑ Label points: SG Symbol

Structure selection:

Structure: 94: 5x5 (a)

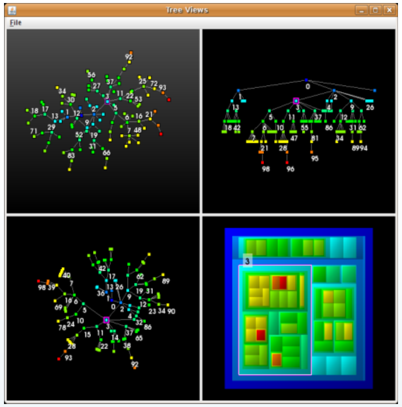
Optimized: 107 | Running: 5 | Failures: 0

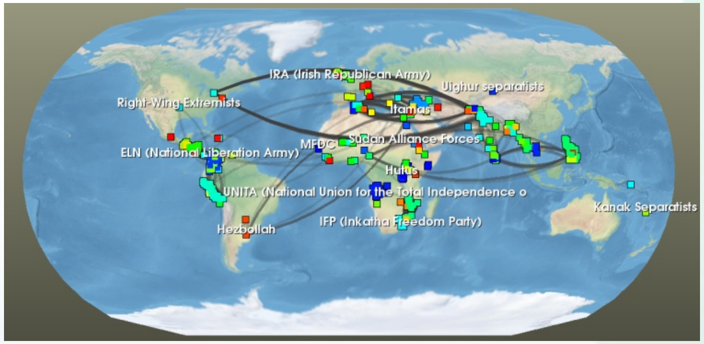
2D: Graphs and Charts

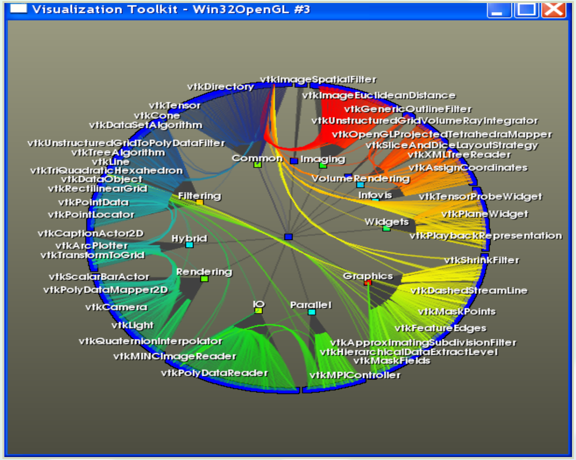


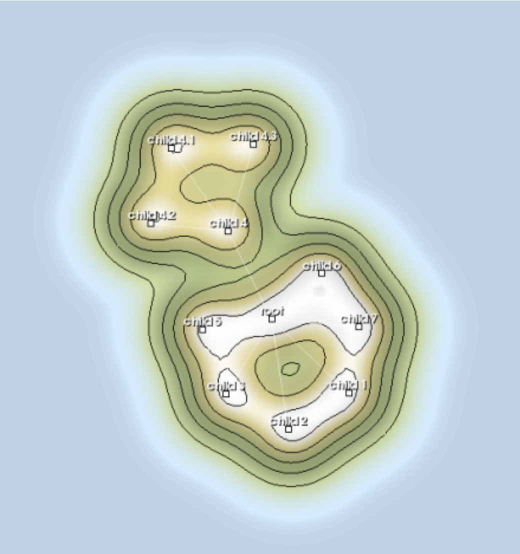
Kitware

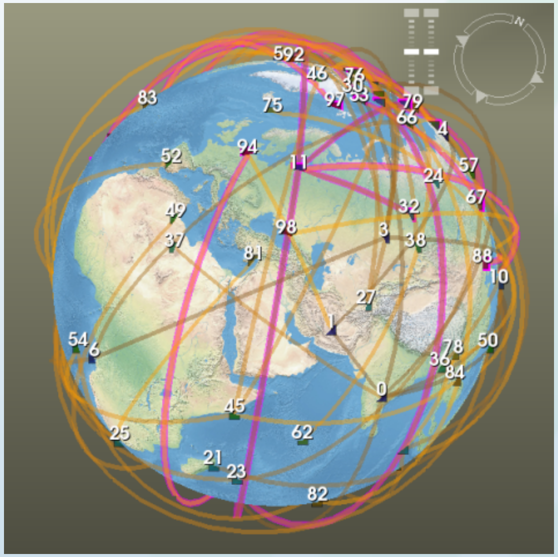
Informatics

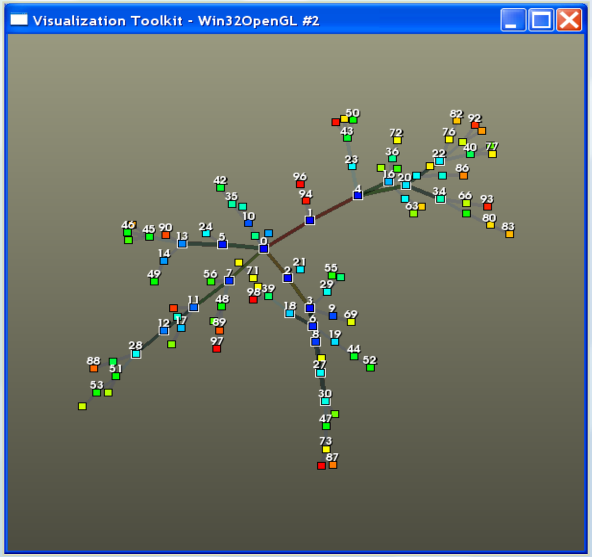



IRA (Irish Republican Army)
Uighur separatists
Right-Wing Extremists
Hamas
MFDC
Sudan Alliance Forces
ELN (National Liberation Army)
Hutus
UNITA (National Union for the Total Independence o
Kanak Separatists
IFP (Inkatha Freedom Party)
Hezbollah

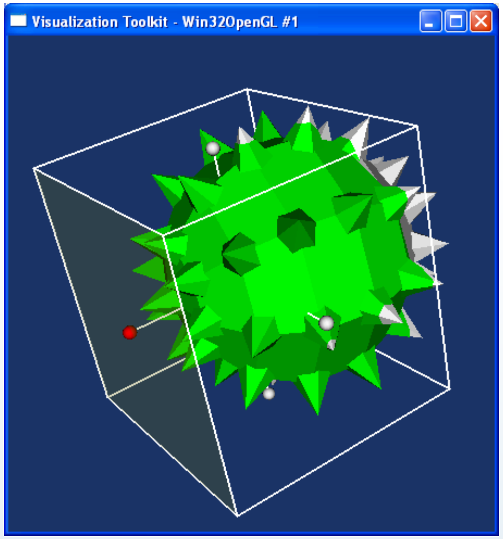


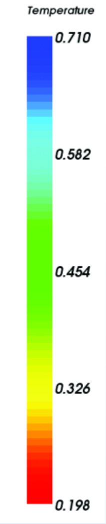


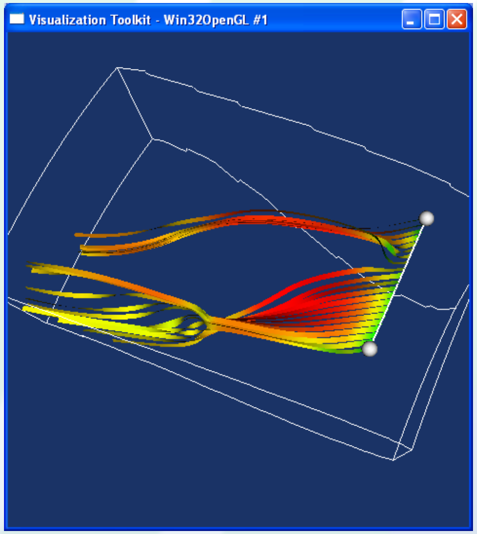


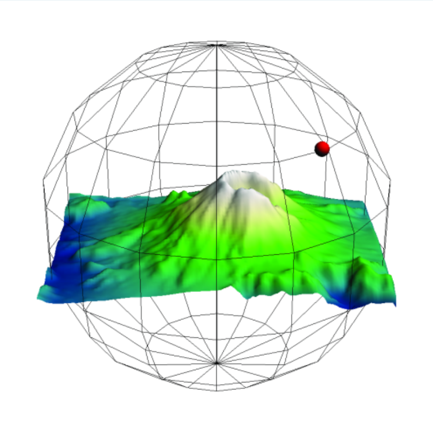


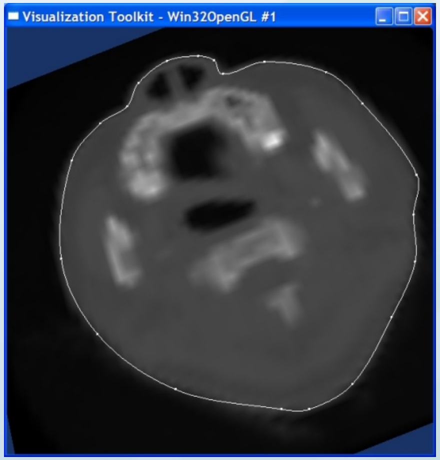
3D Interaction Widgets

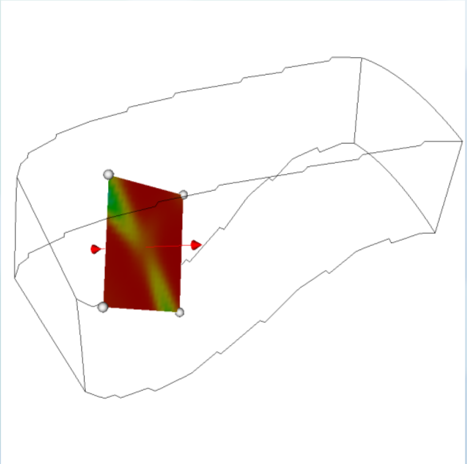












VTK: The Toolkit

- Collection of C++ libraries
  - Leveraged by many applications
  - Divided into logical areas, e.g.
    - Filtering – data processing in visualization pipeline
    - InfoVis – informatics visualization
    - Widgets – 3D interaction widgets
    - VolumeRendering – 3D volume rendering
- Cross platform, using OpenGL
- Wrapped in Python, Tcl and Java

VTK Development Team

- From Ohloh: **Very large, active development team:** Over the past twelve months, **100 developers** contributed new code to VTK. This is one of the largest open-source teams in the world, and is in the **top 2%** of all project teams on Ohloh.

PITTSBURGH SUPERCOMPUTING CENTER

CSCS Swiss National Supercomputing Centre

Los Alamos NATIONAL LABORATORY EST. 1943

CD-adapco

eDF

ARL U.S. Army Research Laboratory

GE

TU Delft Delft University of Technology

Kitware

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Sandia National Laboratories

INDIANA UNIVERSITY

SCI www.sci.utah.edu

GEORGETOWN UNIVERSITY 1789

cea

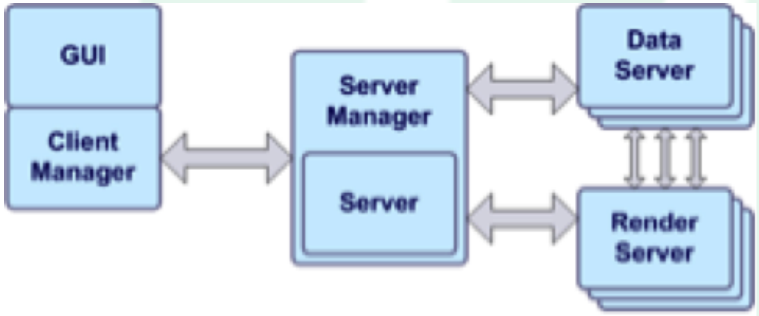
JOHNS HOPKINS UNIVERSITY

MIT Massachusetts Institute of Technology

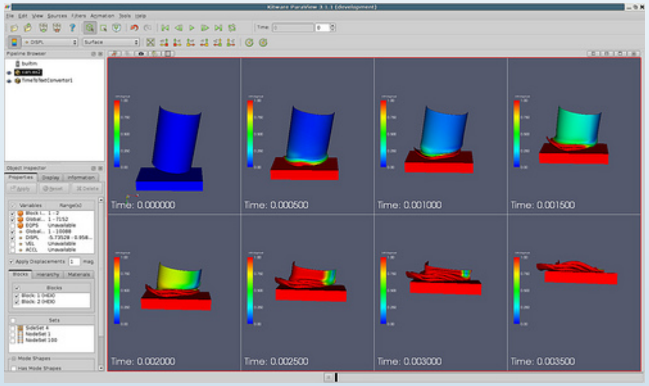
Rensselaer POLYTECHNIC INSTITUTE 1824

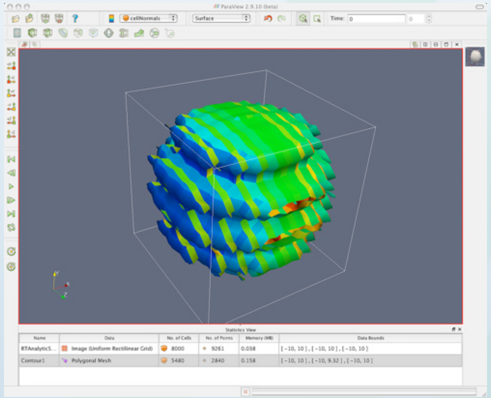
and many others...

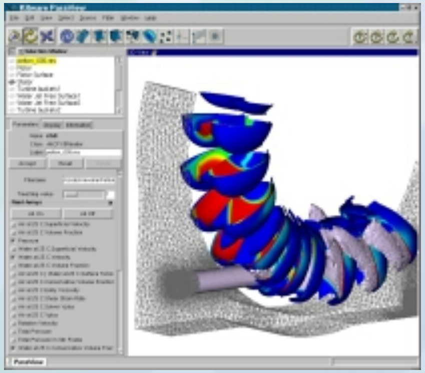
ParaView



- Parallel visualization application
- Open source, BSD licensed
- Turn-key application wrapper around VTK
- Parallel data processing **and** rendering







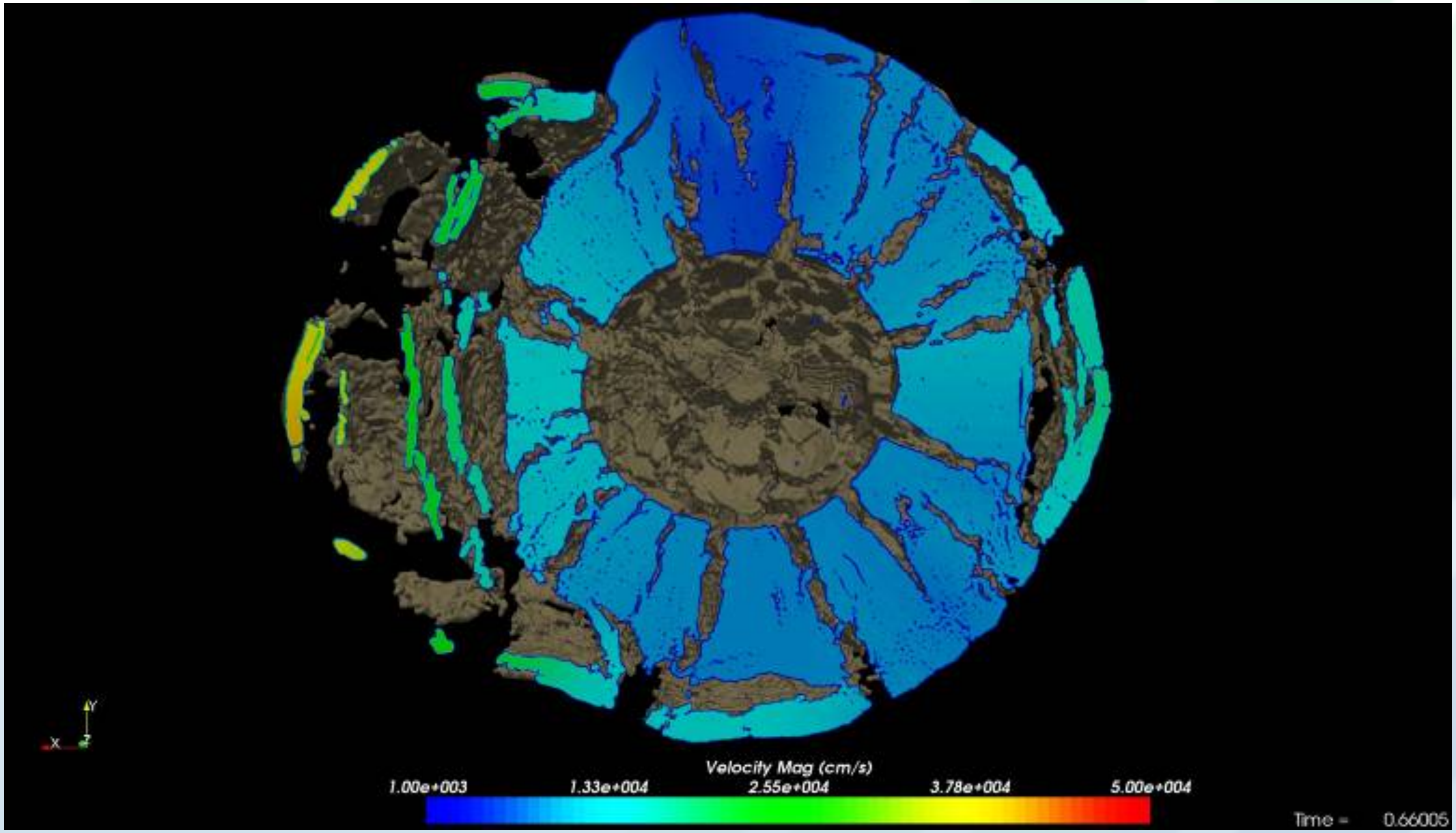
Large Data Visualization

- BlueGene/L at LLNL
  - 65,536 compute nodes (32 bit PPC)
  - 1,024 I/O nodes (32 bit PPC)
  - 512 MB of RAM per node

- Sandia Red Storm
  - 12,960 compute nodes (AMD Opteron dual)
  - 640 service and I/O nodes
  - 40 TB of DDR RAM per node

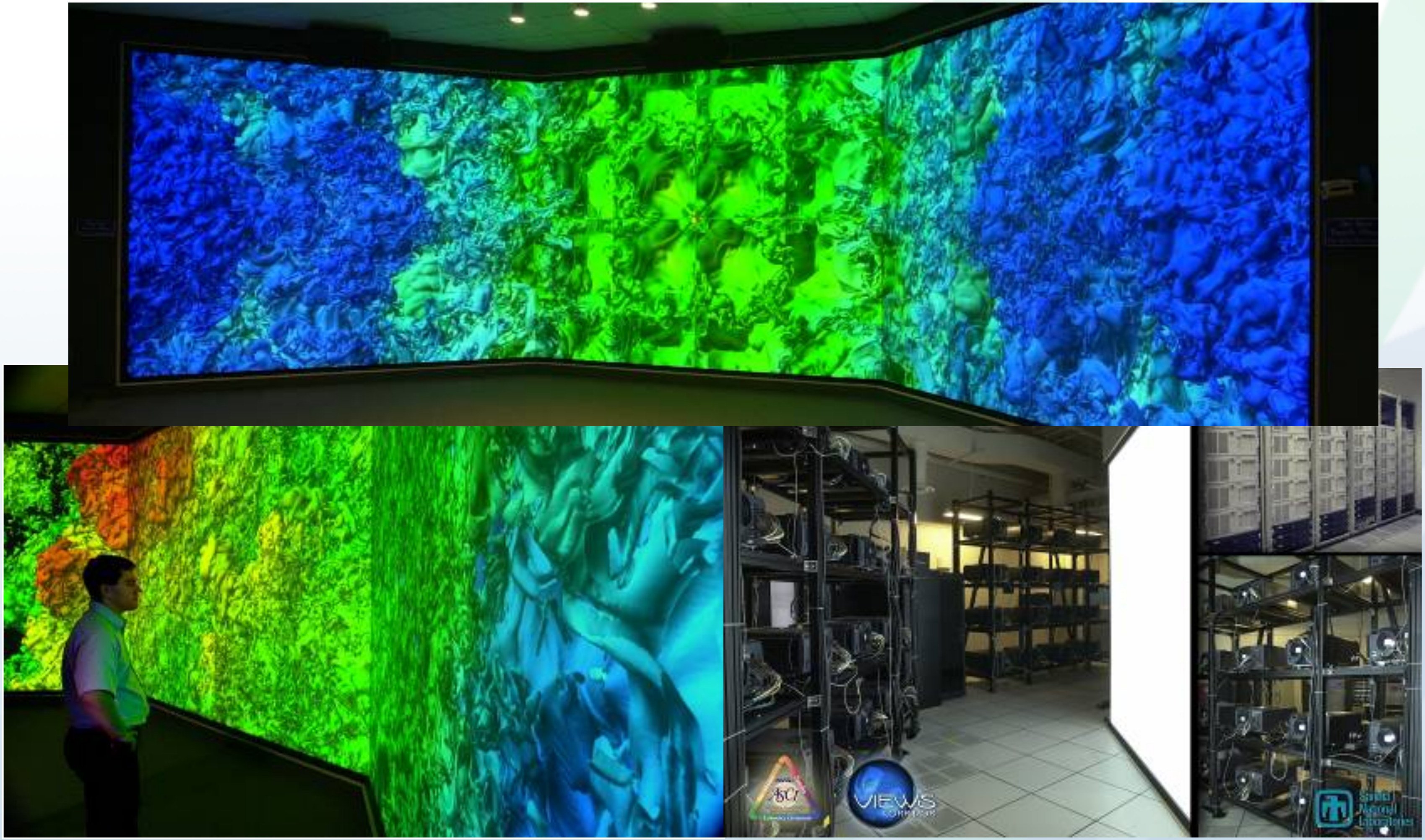
Kitware

1 Billion Cell Asteroid Simulation

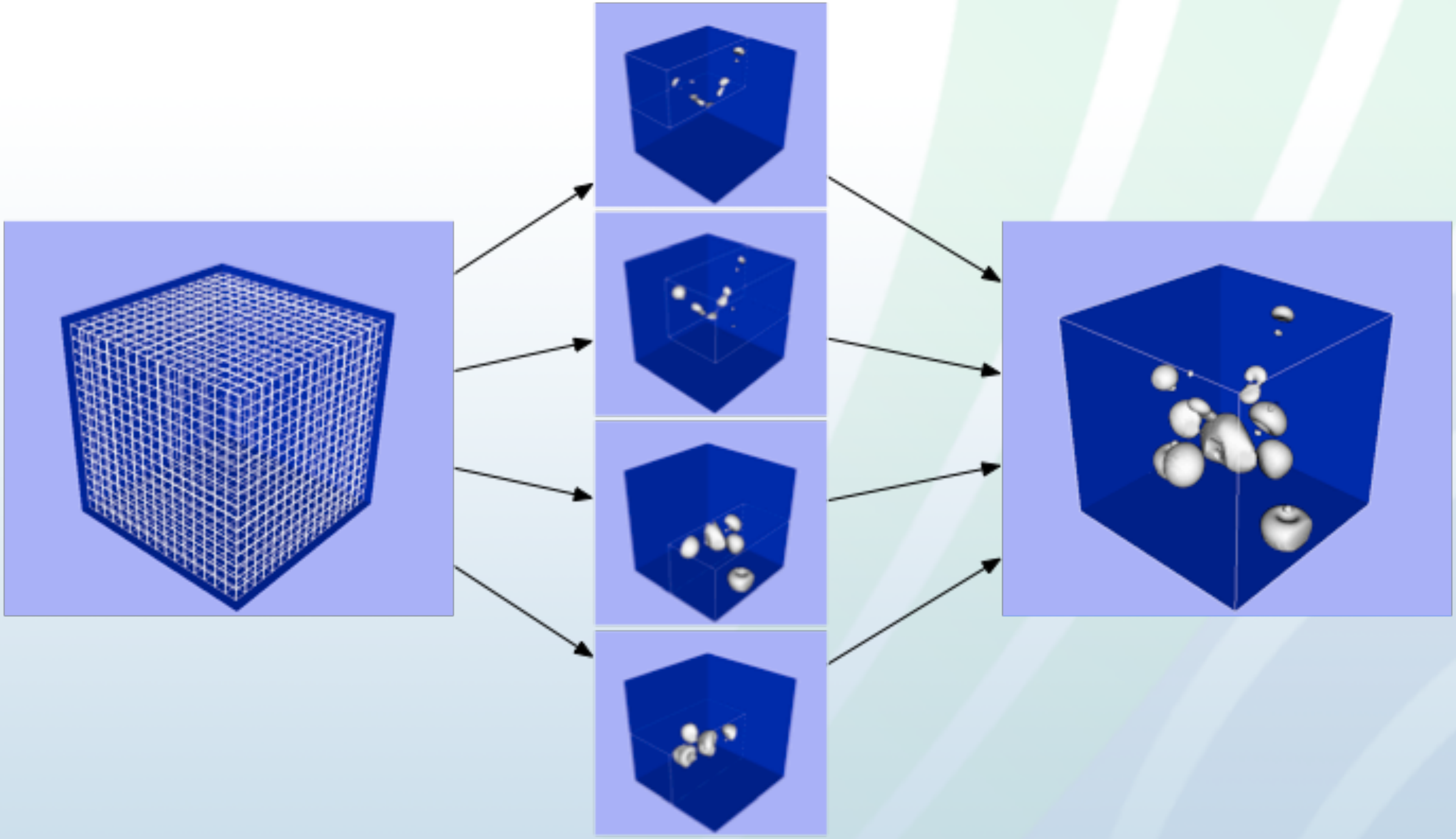


Kitware

Tiled Displays

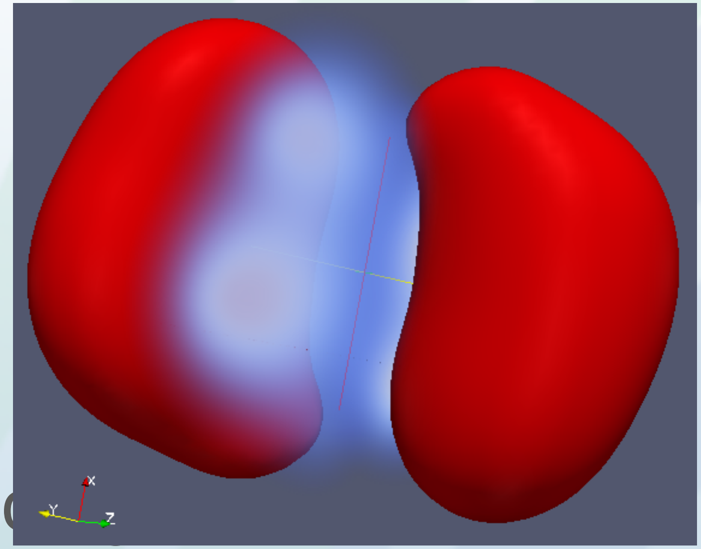


Parallel Processing/Rendering



3D Chemistry Visualization

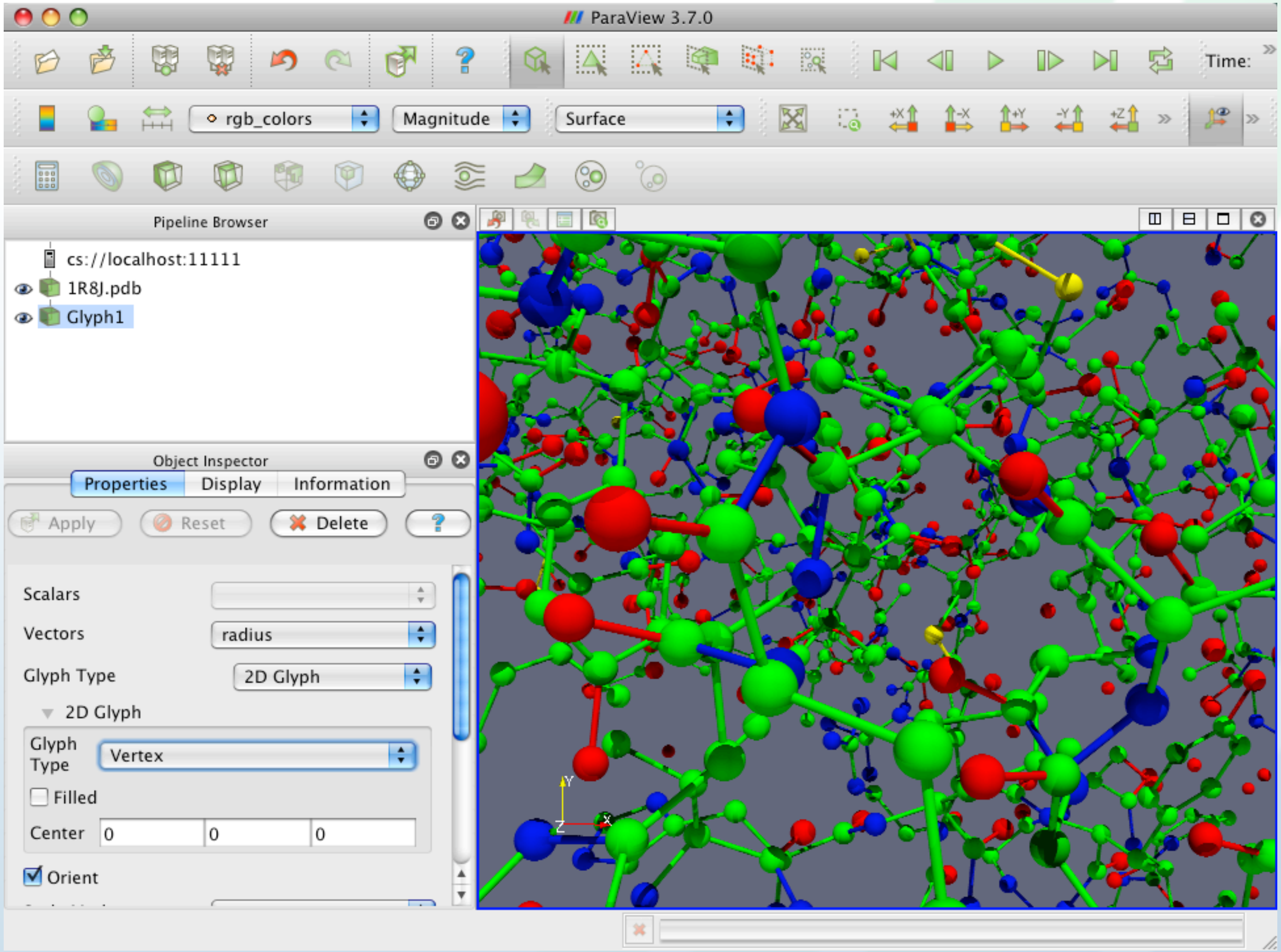
- Some existing features specific to chemistry
  - Gaussian cube, PDB, and a few others
- Excellent handling of volumetric data:
  - Marching cubes
  - Volume rendering
  - Contouring
- Advanced rendering:
  - Point sprites
  - Manta – real time ray trac



Titan: VTK and Informatics

- Led by Sandia National Laboratories
- Substantial expansion of VTK:
  - Informatics & analysis
- Actively developed, growing feature set
- Improved 2D rendering and API
- Database connectivity, client-server, pipeline based approach
- Uses web technologies such as ProtoViz
- Scalable, interactive infoviz

Manta: Real Time Ray Tracing



rgb_colors | Magnitude | Surface

cs://localhost:11111
1R8J.pdb
Glyph1

Scalars
Vectors: radius
Glyph Type: 2D Glyph
2D Glyph
Glyph Type: Vertex
☐ Filled
Center: 0 0 0
☑ Orient

New Frontiers

- New work porting VTK
  - Use C++ as the common core
    - iOS port in the early stages
    - Android port
  - Use OpenGL ES 2.0 – new rendering code
- Also ParaViewWeb – delivering over web
  - Use image delivery and rendering on server
  - Also using WebGL for rendering (optionally)

Future Directions

- VTK modularization (in progress)
  - Developing more agile build systems
  - Automating more with CMake
- Using Git more fully to improve stability
  - Use of master and next
  - Topic branches - merge when ready
- Code review using Gerrit
  - Integration with continuous integration
  - Test before merge